Proceedings of the 6th World Congress on Electrical Engineering and Computer Systems and Sciences (EECSS'20)
~~Prague, Czech Republic~~ Virtual Conference – August, 2020
Paper No. ICBES 120
DOI: 10.11159/icbes20.120

Multi-Dataset Training for Skin Lesion Classification on Multimodal and Multitask Deep Learning

Tudor Nedelcu, Maria Vasconcelos, André Carreiro
Fraunhofer Portugal AICOS
Rua Alfredo Allen 455,4200-135 Porto, Portugal
tudor.nedelcu@fraunhofer.pt; maria.vasconcelos@fraunhofer.pt
andre.carreiro@fraunhofer.pt

Abstract – According to the World Health Organization, skin cancer represents approximately one third of every diagnosed cancer, reaching over 3 million cases over the world, annually. Similar to other types of cancer, though, early diagnosis is key for a good outcome, and computer-aided diagnosis has shown great promise in such task. In this paper we improve the results of previous work on skin lesion diagnosis by using a deep convolutional neural network trained on multimodal data, namely macroscopic and dermoscopic image and metadata. For a deep learning approach is important to have a large number of samples, which EDRA dataset does not present. We have improved the results of previous work in the field of multimodal and multitasking for skin lesion classification by performing transfer learning using similar datasets, which are predicting different skin conditions. By pre-training on datasets which belong to a similar domain, the network learns useful features which enhances the performances of the network.

Keywords**:** Skin cancer, Transfer learning, Multimodal, Multitasking

1. Introduction

Skin cancer is the most common malignancy in fair-skinned populations, and the incidences of melanoma and non-melanoma skin cancers are rising, resulting in high economic costs [1]. Early melanoma diagnosis is crucial for improving the patient condition. Skin cancer is initially diagnosed by a visual inspection from a clinical expert. The initial step consists of a clinical screening, followed by a dermoscopic analysis (non-invasive in-vivo imaging technique, uncovers detailed morphological and visual properties of pigmented lesions), and a biopsy and histopathological examination if necessary. Because some patients are unable to go to the doctor for a visit (e.g. pandemic crisis such as COVID-19, living in remote areas, reduced mobility), tele-dermatology has gained more popularity.

Nowadays, computer-assisted medical diagnosis of skin conditions has undergone major advances. Tele-dermatology has improved with respect to the image acquisition devices and the development of algorithms to process the information. Furthermore, due to medical data storage and the advances in Machine Learning, the development of automatic skin lesion classification has grown considerably [2].

Automated classification of skin lesions from digital images is a challenging task due to the variations of acquired images and to the complexity of this problem. Automated lesion classification can both support physicians in their daily clinical routine and enable fast and cheap access to lifesaving diagnoses, even outside the hospital, through installation of apps on mobile devices [3].

So far most automatic classification methods only consider one image type, although the clinical decision takes into consideration both clinical and dermoscopic images, as well as clinical information from the patient. Codella [4] reported that for experienced dermatologists, the accuracy in diagnosing pigmented skin lesions is improved when a dermatoscope is used. The usage of such methods provides useful inputs to the dermatologist, and even non-specialists might be able to monitor and follow-up suspected skin cancer cases [5].

This work aims to research on multimodal and multitasking for skin lesion classification and explore a transfer learning approach to improve the network performance. This paper is structured as follows: Section 1 presents the motivation and objectives of this work; Section 2 presents the related work along with a description of the datasets used; Section 3 details the proposed methodology used; in Section 4 the results and discussion are presented; Section 5 highlights the main conclusions of this study and points out possible directions for future work.

2. Related work

One of the most significant milestones regarding automatic skin lesion classification is represented by the work of Esteva [6]. The authors have collected 129450 macroscopic images consisting of 2032 diseases. A Deep Neural Network (DNN) was trained on an Inception-V3 architecture using transfer learning. The weights of Inception-V3 pre-trained on ImageNet [7] were used to enhance the learning process. The prediction performances were tested against 21 board-certified dermatologists on biopsy-proven clinical images with two critical binary classification use cases: keratinocyte carcinomas against benign seborrheic keratoses; and malignant melanomas versus benign nevi. The authors used a hierarchical partitioning algorithm using a taxonomy tree for data balancing.

Data available for training the model is a major factor for achieving high accuracy and generalization for unseen data. Han [8] has merged different public datasets with a proprietary dataset to gather over 20.000 samples of macroscopic images, with 12 classified diseases. The ResNet [9] network architecture with weights pre-trained on ImageNet was used for this approach. Transfer learning was performed using the freezing layer approach, where the weights of lower-level layers of the network are not modified, in order to preserve the basic feature extraction from images.

Although the methods described above are designed for macroscopic images, if dermoscopic images are available, the DNN's can be easily trained for dermoscopic images with minor alterations. The interest in skin lesion diagnosis using dermoscopic images increased after the ISIC dataset challenge was introduced [10], alongside with the benchmark for evaluation. The best performing approach was obtained by using an ensemble of DNNs and enhancing the number of samples by merging other datasets.

There are also methods developed upon both types of images and including additional metadata. In [11] the model for diagnosis prediction is generated by joining two InceptionV3 networks with pre-trained weights on ImageNet [7]. All three modalities are used as input to predict the diagnosis and the 7-point checklist [12]. Since the 7-point checklist and the diagnosis are related tasks, the results are improved since the multitasking prediction makes the results more robust [13]. Furthermore, the model performance is increased by combining the outputs of the complementary modalities. During training all the modalities are available, whereas for the inference stage one or a specific combination of modalities can be used.

Yap [14] has also worked on multimodal data for skin lesion classification. The ResNet50 network with weights pre-trained on ImageNet [7] was used to reduce the overfitting for a relatively small database (2917 cases). Despite more images were available initially, some samples were removed from the dataset during the data curation. For example, images of poor quality or images where a part of the body is identifiable were removed. Therefore, the risk of biased data towards classification of certain areas of the body was reduced. It is worth mentioning that in this study [14] is noticed that merging the two image modalities improved the results considerably, whereas the metadata input had a marginal impact.

2.1. Datasets

Although there are several datasets available for skin cancer analysis, it is not viable to merge all of them into a global one. They cannot be trained together using all the available inputs, due to the variations amongst the metadata used, types of images (dermoscopic and macroscopic), or even the predicted output (classes). In this work we select the publicly available datasets which have macroscopic or dermoscopic images and the diagnostic labels are related.

The EDRA [11] dataset contains samples of various modalities (macroscopic and dermoscopic images, and metadata). This dataset consists of 1011 images for each image modality (total of 2022 images). Alongside the images, relevant information as patient metadata and the 7-point checklist is provided (Table 1b). The diagnosis consists of a basal cell carcinoma - BCC, nevus - NV (blue, clark, combined, congenital, dermal, recurrent and reed nevus), melanoma – MEL (in situ, less than 0.76mm, between 0.76-1.5mm, metastasis), miscellaneous – MISC (dermatofibroma, lentigo, melanosis, miscellaneous, vascular lesion), and seborrheic keratosis - SK. Regarding the 7-point checklist, there is the annotation of the presence\absence of the feature, and if it is regular, irregular or atypical. Although there are various types of diagnosis, and 7-point checklist features, they are grouped into the main classes

Table 1: Datasets used and the 7-points checklist data from EDRA dataset.

a) Dataset	Diagnostic	samples
EDRA	Basal cell carcinoma (BCC)	42
	Nevus (NEV)	575
	Melanoma (MEL)	252
	Miscellaneous (MISC)	97
	Seborrheic keratosis (SK)	45
ISIC 2019	Melanoma	4522
	Melanocytic nevus	12875
	Basal cell carcinoma	3323
	Actinic keratosis	867
	Benign keratosis	2624
	Dermatofibroma	239
	Vascular lesion	253
	Squamous cell carcinoma	628
Dermofit	Actinic Keratosis	45
	Basal cell carcinoma	239
	Melanocytic nevus	331
	Squamous cell carcinoma	88
	Seborrheic keratosis	257
	Intraepithelial carcinoma	78
	Pyogenic granuloma	24
	Haemangioma	96
	Dermatofibroma	65
	Melanoma	76

b) 7-Point Checklist (EDRA)		
	Type	7-point score
Piment Network (PN)	Absent (ABS)	0
	Typical TYP	0
	Atypical ATP	2
Blue Whitish Veil (BWV)	Absent (ABS)	0
	Present PRS	2
Vascular Structures (VS)	Absent (ABS)	0
	Regular (REG)	0
	Irregular (IR)	2
Pigmentation (PIG)	Absent (ABS)	0
	Regular (REG)	0
	Irregular (IR)	1
Streaks (STR)	Absent (ABS)	0
	Regular (REG)	0
	Irregular (IR)	1
Dots and Globules (DaG)	Absent (ABS)	0
	Regular (REG)	0
	Irregular (IR)	1
Regression Structures (RS)	Absent (ABS)	0
	Present (PRS)	1

(pigment network – PN, blue whitish veil – BWV, vascular structures – VS, pigmentation – PIG, streaks – STR, dots and globules – DaG, and regression structures - RS). The available metadata is represented by the location of the skin lesion on the body, its elevation, and the gender of the patient. For further comparison with other methods, the authors have also proposed a splitting of the dataset into 413 samples for training, 203 for validation, and 395 for testing.

The ISIC 2019 challenge dataset is composed of several datasets (Ham1000 [15], BCN_20000 [16], and MSK [4]). All the images of ISIC 2019 are dermoscopic ones. The 25331 images have associated metadata and are labelled regarding the following skin conditions: melanoma, melanocytic nevus, basal cell carcinoma, actinic keratosis, benign keratosis, dermatofibroma, vascular lesion, and squamous cell carcinoma (Table 1a).

On the other hand, the Dermofit [17] dataset consists of macroscopic images only. The Dermofit Image Library is a collection of 1,300 focal high-quality skin lesion images collected under standardised conditions with internal colour standards. The lesions span across ten different classes including melanomas, seborrheic keratosis and basal cell carcinomas (Table 1a). Each image has a gold standard diagnosis based on expert opinion (including dermatologists and dermato-pathologists). Images consist of a snapshot of the lesion surrounded by some normal skin. A binary segmentation mask that denotes the lesion area is included with each lesion. The categories of lesions are actinic keratosis, basal cell carcinoma, melanocytic nevus, seborrheic keratosis, intraepithelial carcinoma, pyogenic granuloma, haemangioma, dermatofibroma, and malignant melanoma.

3. Proposed Method

Since the EDRA dataset has a small number of samples for a very complex task, we decided to enhance the performance of a DNN model by using domain adaptation. The other available datasets present a significant variation amongst them (number of samples, prediction of various skin diseases, different metadata, various modalities, etc.), but on the other hand the prediction of the outputs is in the same domain as the EDRA dataset. In this work we have performed transfer learning by training on two databases from a similar domain (Fig. 1).

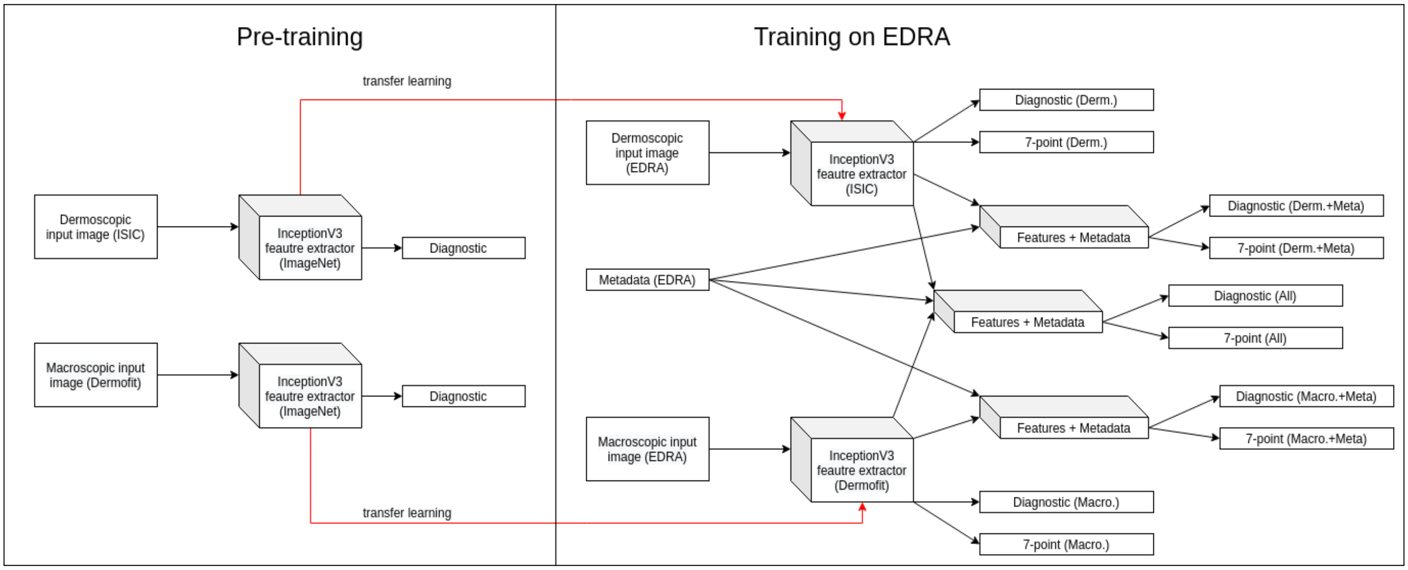

Fig. 1: Proposed pipeline for transfer learning of multimodal approach on multitask learning.

3.1. Multimodal/multitasking network

The DNN model used in this work is the one proposed by Kawahara [11]. A multimodal network is deployed by using two InceptionV3 models pre-trained on ImageNet [7] to predict the diagnosis and the 7-point checklist. The two networks are trained jointly using late fusion and metadata. There are 5 modalities proposed for training the network: macroscopic images, dermoscopic images, metadata and macroscopic images, metadata and dermoscopic images, and a combination of all three (macroscopic and dermoscopic images and metadata). The network parameters are updated with respect to the loss of all the modalities:

$$L(x, y; \theta) = \sum_{j=1}^{8} l(x, y_j; \theta), \tag{1}$$

where $l(\cdot)$ is the categorical cross-entropy operator, *x* is the input (5 modalities), *y* is the output of the diagnosis and 7-point checklist, and θ is a set of trainable parameters.

The classification layers are removed from the InceptionV3, and a new classification layer is added for each modality. Instead of using a dense layer, the global average pooling followed by a *softmax* layer is adopted. This classification layer outperforms the dense layer since it reduces the overfitting during training. We have experimented with a dense layer approach, and we concluded that the global average pooling has a major benefit for network performance.

3.2. Pre-training

Since the proposed network consists of two InceptionV3 networks (one network for each type of image), we have pre-trained these models accordingly. For the dermoscopic images the ISIC2019 dataset [4], [15], [16] was used, and for the macroscopic images the Dermofit dataset [17] was used. The models were trained separately using the same hyperparameters as [11]. For all the models a learning rate of 1e-3 was used, with a decay of 1e-4, and the SGD optimizer. For both datasets, the data was split for training and validation by selecting 90% and 10% of the data, respectively. To ensure we have a uniform distribution of classes, we have selected the samples from datasets accordingly. To deal with imbalanced data, we have used augmentation by up-sampling. The data augmentation consists of several image processing operations like flipping, rotating, zooming and height and width shifts.

The freezing layers approach was used for training the dermoscopic and macroscopic models, to enhance the learning process. Similar to Kawahara [11], the first two blocks were excluded from learning. Since the ISIC2019 dataset has more samples, we used only 3 epochs for each InceptionV3 block, and for the model trained on macroscopic data 5 epochs for each block were used.

After the training process is finished, the classification layers of both networks are removed and the two InceptionV3 networks are used to extract useful features related with skin lesions. We have implemented a modified version of Kawahara's network, and loaded the weights generated by training the two InceptionV3 models on ISIC2019 and Dermofit datasets. Our multimodal network differs from Kawahara's approach through the addition of the metadata after the prediction of macroscopic and dermoscopic modalities.

4. Results

Because of the scarcity of the data is very important to be able to predict a diagnosis regardless of missing data. In Table 2 the accuracy results of our proposed training procedure is depicted. There is an increase of 5% in accuracy regarding the prediction of the diagnosis when compared to Kawahara's method [11]. Although the metadata might provide useful information during training due to the aggregated loss function, the combination of image and metadata reveals marginal or no improvements over the single image modalities. This behaviour is noticed in the literature [7], where the addition of metadata had a marginal improvement. The lowest accuracy is achieved for macroscopic images, because the clinical images are providing less details compared with the dermoscopic ones. Although the metadata does not improve the accuracy of the diagnostic, the classification of the 7-points checklist is improved. There is no direct improvement of the diagnostic by using the metadata, but during training we suspect that the network is learning useful features which allow a better generalization. We can notice that there is an overall improvement over the direct classification for the diagnostic. For prediction of the melanoma from the 7-point checklist criteria, the method proposed by Kawahara is obtaining better results. One possible explanation for this is related to the available data for pre-training the networks (dermoscopic and macroscopic), where only the diagnostic was used as a target.

Table 2: Accuracy for each of the 7-point checklist criteria and diagnosis.

Modality	**BWV**	**DaG**	**PIG**	**PN**	**RS**	**STR**	**VS**	**DIAG**
Macro	76.5	47.1	52.4	54.2	68.4	58.5	62.5	61.8
Derma	85.6	54.7	63.5	65.1	**78.7**	72.9	80.0	78.0
Macro-Meta	77.2	48.4	51.1	55.2	68.4	59.0	64.1	61.8
Derma-Meta	86.3	58.7	63.8	65.6	**78.7**	73.4	80.5	77.0
All	85.6	57.5	64.6	65.1	77.2	71.1	80.3	75.7
x_combine	85.3	57.5	64.6	64.6	78.5	71.9	**81.3**	**79.2**
x_combine [11]	**87.1**	**60.0**	**66.1**	**70.9**	77.2	**74.2**	79.7	74.2

In Table 3 are the results of the 7-point checklist prediction regarding the sensitivity, specificity and precision. Since melanoma is estimated by a linear combination of the 7-point checklist [12], we have evaluated this prediction also. To compute the 7-points checklist score for melanoma, the look-up table is used (Table 1b). The obtained score is truncated using a threshold (1, or 3) to predict if the skin lesion is malignant:

$$S = \sum_{j=1}^{7} y_j w, \tag{2}$$

$$M = \begin{cases} melanoma, & S \geq t \\ not\ melanoma, & S < t \end{cases}, \tag{3}$$

where S is the 7-points score, y_j is the prediction of the 7-points labels, and w is the associated weight. The sample is classified as melanoma if the 7-points score is above a certain threshold t

The *x_combine* represents the result of mixing several modalities (average of derma, derma-meta, and all modalities). By combining several outputs, the results are improving. This is similar with an ensemble of networks, where the majority of the networks will predict a good result usually.

Table 3: Results of diagnostic category and melanoma using the 7-point checklist point scores.

Modality	**Metric**	**DIAG**					**Avrg.**	**MEL-7pt**	
		BCC	**NEV**	**MEL**	**MISC**	**SK**		**t=1**	**t=3**
x-combine	Sens.	56.2	93.2	67.3	60.0	42.1	**63.8**	93.1	66.3
	Spec.	98.4	73.9	94.6	97.2	98.9	**92.6**	**43.9**	84.0
	Prec.	60.0	81.6	81.0	70.6	66.7	**72.0**	36.3	58.8
x_combine[11]	Sens.	62.5	88.6	61.4	47.5	42.1	60.4	**96.0**	**69.3**
	Spec.	97.9	71.6	88.8	97.5	99.5	91.0	36.1	**77.6**
	Prec.	55.6	79.5	65.3	67.9	80.0	69.6	**34.0**	51.5

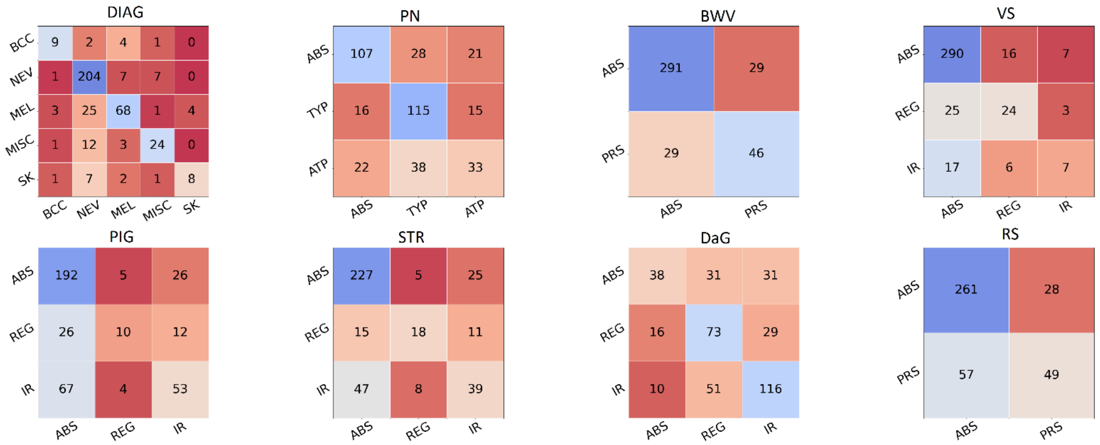

Fig. 2: Confusion matrices for diagnosis and the 7-point prediction of the test set evaluation. The *x*-axis indicates the model's prediction, while the *y*-axis indicates the ground-truth. Colours indicate the percentage of each label for each entry, normalized by the total number of true labels.

Regarding the distribution of the predicted diagnosis and the 7-point checklist, in Fig. 2 we can observe how well the model is performing with respect to each class. Although the training of the network was performed aiming to mitigate the use of an imbalanced dataset, is noticeable a slight bias towards certain classes. For the diagnostic prediction, we noticed that the model is biased towards Nevus, since most of the false negative cases are miss-classified as Nevus (except BCC).

This occurs because the dataset has a few samples, where most of the images are skin lesions with a Nevus condition. The results can be further improved by considering the one vs. all approach when fine-tuning the model.

5. Conclusion

In this work it was investigated the transfer learning for training a multimodal/multitasking network for skin lesion classification. This work was motivated by the scarcity of the available data (EDRA dataset) and by the inconsistency of the available datasets (prediction of different outputs, the metadata is not similar, a single type of image available, etc.). Two networks were trained separately on specific datasets (ISIC for dermoscopic images and Dermofit for macroscopic ones), and the knowledge was transferred to a multimodal network designed to predict the diagnostic and the 7-points checklist using EDRA dataset. The results show that transfer learning can be used to further increase the performance of a DNN framework, by training on samples related to a specific domain (skin conditions). Although the metadata is available for the training of the networks, no significant improvement is observed, even though it can enhance the learning process

Regarding the performances achieved on EDRA dataset, we conclude that this dataset is very challenging, having a limited number of samples, with some of the images that are of low quality. Furthermore, some body parts are identifiable in some images and the model might become biased for a specific class.

As future work, we plan to mitigate the bias of the network by using additional layers to differentiate from classes which have a similar output. E.g. a classification layer to differentiate between Melanoma and Nevus, since 25 out of 33 false negative predictions of Melanoma are classified as Nevus (Fig. 2). Another approach to mitigate the reduced number of samples is by using smart augmentation (GAN, CGAN, WGAN, etc.). Furthermore, pre-training on other datasets which have the 7-point checklist or other additional information, which will be suitable for multitask learning can be also considered as future work.

Acknowledgements

This work was done under the scope of project “DERM.AI: Usage of Artificial Intelligence to Power Teledermatological Screening” and supported by national funds through ‘FCT—Foundation for Science and Technology, I.P.’, with reference DSAIPA/AI/0031/2018.

References

[1] Ali, Abder-Rahman A, and Thomas M Deserno. “A Systematic Review of Automated Melanoma Detection in Dermatoscopic Images and Its Ground Truth Data.” In *Medical Imaging 2012: Image Perception, Observer Performance, and Technology Assessment*, 8318:83181I. International Society for Optics and Photonics, 2012.

[2] Chan, S., Reddy, V., Myers, B., Thibodeaux, Q., Brownstone, N., & Liao, W. (2020). Machine Learning in Dermatology: Current Applications, Opportunities, and Limitations. Dermatology and Therapy, 1-22.

[3] De Carvalho, T.M.; Noels, E.; Wakkee, M.; Udrea, A.; Nijsten, T. Development of Smartphone Apps for Skin Cancer Risk Assessment: Progress and Promise. JMIR Dermatol. 2019, 2, e13376..

[4] Noel C. F. Codella, David Gutman, M. Emre Celebi, Brian Helba, Michael A. Marchetti, Stephen W. Dusza, Aadi Kalloo, Konstantinos Liopyris, Nabin Mishra, Harald Kittler, Allan Halpern: “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)”, 2017; arXiv:1710.05006

[5] Ma, Zhen, João Manuel RS Tavares, and others. “A Review of the Quantification and Classification of Pigmented Skin Lesions: From Dedicated to Hand-Held Devices.” *Journal of Medical Systems* 39, no. 11 (2015): 177.

[6] Esteva, Andre, Brett Kuprel, Roberto A Novoa, Justin Ko, Susan M Swetter, Helen M Blau, and Sebastian Thrun “Dermatologist-Level Classification of Skin Cancer with Deep Neural Networks.” Nature 542, no. 7639 (2017): 115– 118.

[7] Russakovsky, Olga, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein. “Imagenet Large Scale Visual Recognition Challenge.” International

Journal of Computer Vision 115, no. 3 (2015): 211–252.

[8] Han, Seung Seog, Woohyung Lim, Myoung Shin Kim, Ilwoo Park, Gyeong Hun Park, and Sung Eun Chang. "Interpretation of the Outputs of a Deep Learning Model Trained with a Skin Cancer Dataset." The Journal of Investigative Dermatology 138, no. 10 (2018): 2275.

[9] He, Kaiming, 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, 770–778, 2016.

[10] Gutman, D., Codella, N. C., Celebi, E., Helba, B., Marchetti, M., Mishra, N., & Halpern, A. (2016). Skin lesion analysis toward melanoma detection: A challenge at the international symposium on biomedical imaging (ISBI) 2016, hosted by the international skin imaging collaboration (ISIC). arXiv preprint arXiv:1605.01397.

[11] Kawahara, Jeremy, Sara Daneshvar, Giuseppe Argenziano, and Ghassan Hamarneh. "Seven-Point Checklist and Skin Lesion Classification Using Multitask Multimodal Neural Nets." IEEE Journal of Biomedical and Health Informatics 23, no. 2 (2018): 538–546.

[12] Argenziano, Giuseppe, Gabriella Fabbrocini, Paolo Carli, Vincenzo De Giorgi, Elena Sammarco, and Mario Delfino. "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 134, no. 12 (1998): 1563–1570.

[13] Seltzer, Michael L, and Jasha Droppo. "Multi-Task Learning in Deep Neural Networks for Improved Phoneme Recognition." In 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, 6965–6969. IEEE, 2013.

[14] Yap, Jordan, William Yolland, and Philipp Tschandl. "Multimodal Skin Lesion Classification Using Deep Learning." *Experimental Dermatology* 27, no. 11 (2018): 1261–1267.

[15] Tschandl, Philipp, Cliff Rosendahl, and Harald Kittler. "The HAM10000 Dataset, a Large Collection of Multi-Source Dermatoscopic Images of Common Pigmented Skin Lesions." Scientific Data 5 (2018): 180161.

[16] Marc Combalia, Noel C. F. Codella, Veronica Rotemberg, Brian Helba, Veronica Vilaplana, Ofer Reiter, Allan C. Halpern, Susana Puig, Josep Malvehy: "BCN20000: Dermoscopic Lesions in the Wild", 2019; arXiv:1908.02288

[17] Ballerini, Lucia, Robert B Fisher, Ben Aldridge, and Jonathan Rees. "A Color and Texture Based Hierarchical K-NN Approach to the Classification of Non-Melanoma Skin Lesions." In Color Medical Image Analysis, 63–86. Springer 2013.