MethodsX 7 (2020) 100864

Contents lists available at ScienceDirect

MethodsX

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

Method Article

Skin lesion classification using ensembles of multi-resolution EfficientNets with meta data ☆

Nils Gessert [a,b,*], Maximilian Nielsen [b,c], Mohsin Shaikh [b,c], René Werner [b,c], Alexander Schlaefer [a,b]

[a] Institute of Medical Technology, Hamburg University of Technology, Hamburg, Germany
[b] DAISYlab, Forschungszentrum Medizintechnik Hamburg, Hamburg, Germany
[c] Institute of Computational Neuroscience, University Medical Center Hamburg-Eppendorf, Hamburg, Germany

A B S T R A C T

In this paper, we describe our method for the ISIC 2019 Skin Lesion Classification Challenge. The challenge comes with two tasks. For task 1, skin lesions have to be classified based on dermoscopic images. For task 2, dermoscopic images and additional patient meta data are used. Our deep learning-based method achieved first place for both tasks. The are several problems we address with our method. First, there is an unknown class in the test set which we cover with a data-driven approach. Second, there is a severe class imbalance that we address with loss balancing. Third, there are images with different resolutions which motivates two different cropping strategies and multi-crop evaluation. Last, there is patient meta data available which we incorporate with a dense neural network branch.

- We address skin lesion classification with an ensemble of deep learning models including EfficientNets, SENet, and ResNeXt WSL, selected by a search strategy.
- We rely on multiple model input resolutions and employ two cropping strategies for training. We counter severe class imbalance with a loss balancing approach.
- We predict an additional, unknown class with a data-driven approach and we make use of patient meta data with an additional input branch.

A R T I C L E I N F O

Method name: Convolutional neural network
Keywords: Deep Learning, Multi-class skin lesion classification, Convolutional neural networks
Article history: Received 27 December 2019; Accepted 9 March 2020; Available online 19 March 2020

☆ **Direct Submission or Co-Submission** *Co-submissions are papers that have been submitted alongside an original research paper accepted for publication by another Elsevier journal*

* Corresponding author at: Institute of Medical Technology, Hamburg University of Technology, Hamburg, Germany.
E-mail address: nils.gessert@tuhh.de (N. Gessert).

https://doi.org/10.1016/j.mex.2020.100864

Specifications Table

Subject Area:	Computer Science
More specific subject area:	Deep learning and skin lesion classification
Method name:	Convolutional neural network
Name and reference of original method:	Not applicable – our method is based on multiple approaches which we cite and detail in the method description
Resource availability:	Public code: https://github.com/ngessert/isic2019 Challenge results: https://challenge2019.isic-archive.com/leaderboard.html Datasets: • https://challenge2019.isic-archive.com/data.html (official) • https://github.com/jeremykawahara/derm7pt (7-point)

Method details

Datasets

The main training dataset contains 25331 dermoscopic images, acquired at multiple sites and with different preprocessing methods applied beforehand. It contains images of the classes melanoma (MEL), melanocytic nevus (NV), basal cell carcinoma (BCC), actinic keratosis (AK), benign keratosis (BKL), dermatofibroma (DF), vascular lesion (VASC) and squamous cell carcinoma (SCC). A part of the training dataset is the HAM10000 dataset which contains images of size 600×450 that were centered and cropped around the lesion. The dataset curators applied histogram corrections to some images [1]. Another dataset, BCN_20000, contains images of size 1024×1024. This dataset is particularly challenging as many images are uncropped and lesions in difficult and uncommon locations are present [2]. Last, the MSK dataset contains images with various sizes.

The dataset also contains meta-information about the patient's age group (in steps of five years), the anatomical site (eight possible sites) and the sex (male/female). The meta data is partially incomplete, i.e., there are missing values for some images.

In addition, we make use of external data. We use the 955 dermoscopic images from the 7-point dataset [3]. Moreover, we use an in-house dataset which consists of 986 images. For the unknown class, we use 353 images obtained from a web search. We include images of healthy skin, angiomas, warts, cysts, and other benign alterations. The key idea is to build a broad class of skin variations that should encourage the model to assign any image that is not part of the eight main classes to the ninth broad pool of skin alterations. We also consider the three types of meta data for our external data, if it is available.

For internal evaluation, we split the main training dataset into five folds. The dataset contains multiple images of the same lesions. Thus, we ensure that all images of the same lesion are in the same fold. We add all our external data to each of the training sets. Note that we do not include any of our images from the unknown class in our evaluation as we do not know whether they accurately represent the actual unknown class. Thus, all our models are trained to predict nine classes but we only evaluate on the known, eight classes.

We use the mean sensitivity for our internal evaluation which is defined as

$$S = \frac{1}{C} \sum_{i=1}^{C} \frac{TP_i}{TP_i + FN_i}$$

where TP are true positives, FN are false negatives and C is the number of classes. The metric is also used for the final challenge ranking.

Image preprocessing

As a first step, we use a cropping strategy to deal with the uncropped images which often show large, black areas. We binarize the images with a very low threshold, such that the entire dermoscopy field of view is set to 1. Then, we find the center of mass and the major and minor axis of an ellipse that has the same second central moments as the inner area. Based on these values we

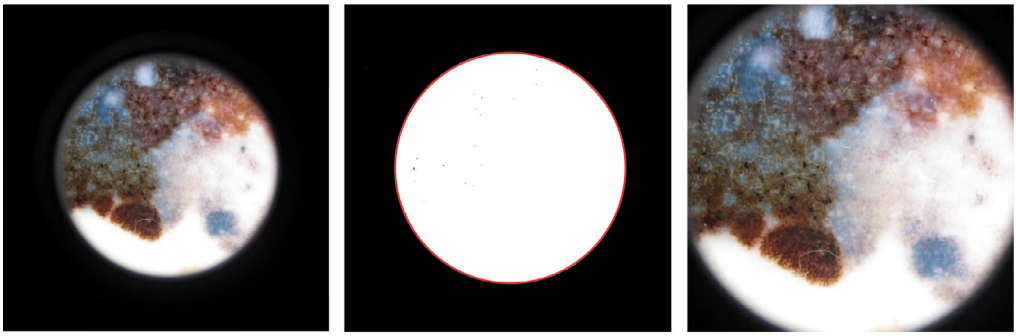

Fig. 1. Cropping strategy for dermoscopic images with a large, black area around the images.

derive a rectangular bounding box for cropping that covers the relevant field of view. The process is illustrated in Fig. 1. We automatically determine the necessity for cropping based on a heuristic that tests whether the mean intensity inside the bounding box is substantially different from the mean intensity outside of the bounding box. Manual inspection showed that the method was robust. In the training set, 6226 were automatically cropped. In the test set, 3864 images were automatically cropped. Next, we apply the Shades of Gray color constancy method with Minkowski norm $p = 6$, following last year's winner [4]. This is particularly important as the datasets used for training differ a lot. Furthermore, we resize the larger images in the datasets. We take the HAM10000 resolution as a reference and resize all images' longer side to 600 pixels while preserving the aspect ratio.

Meta data preprocessing

For task 2, our approach is to use the meta data with a dense (fully-connected) neural network. Thus, we need to encode the data as a feature vector. For the anatomical site and sex, we chose a one-hot encoding. Thus, the anatomical site is represented by eight features where one of those features is one and the others are zero for each lesion. The same applies to sex. In case the value is missing, all features for that property are zero. For age, we use a normal, numerical encoding, i.e. age is represented by a single feature. This makes encoding missing values difficult, as the missingness should not have any meaning (we assume that all values are missing at random). We encode a missing value as -5 as 0 is also part of the training set's value range. To overcome the issue of missing value encoding, we also considered a one-hot encoding for the age groups. However, initial validation experiments should slightly worse performance which is why we continued with the numerical encoding.

Deep learning models

General approach

For task 1, we employ various CNNs for classifying dermoscopic images. For task 2, our deep learning models consist of two parts, a CNN for dermoscopy images and a dense neural network for meta data. The approach is illustrated in Fig. 2. As a first step, we train our CNNs on image data only (task 1). Then, we freeze the CNNs weights and attach the meta data neural network. In the second step, we only train the meta data network's weights and the classification layer. We describe CNN training first, followed by the meta data training.

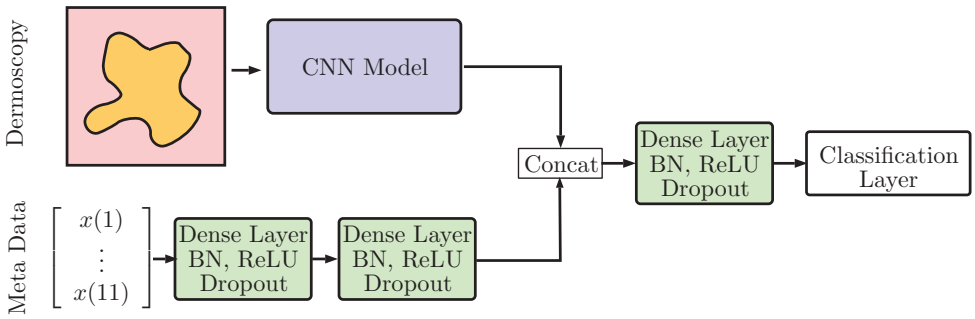

Fig. 2. General approach for combining dermoscopic image processing and meta data processing.

CNN architectures

We largely rely on EfficientNets (EN) [5] that have been pretrained on the ImageNet dataset with the AutoAugment v0 policy [6]. This model family contains eight different models that are structurally similar and follow certain scaling rules for adjustment to larger image sizes. The smallest version B0 uses the standard input size 224×224. Larger versions, up to B7, use increased input size while also scaling up network width (number of feature maps per layer) and network depth (number of layers). We employ EN B0 up to B6. For more variability in our final ensemble, we also include a SENet154 [7] and two ResNext variants pretrained with weakly supervised learning (WSL) on 940 million images [8].

CNN data augmentation

Before feeding the images to the networks, we perform extensive data augmentation. We use random brightness and contrast changes, random flipping, random rotation, random scaling (with appropriate padding/cropping), and random shear. Furthermore, we use CutOut [9] with one hole and a hole size of 16. We tried to apply the AutoAugment v0 policy, however, we did not observe better performance.

CNN input strategy

We follow different input strategies for training that transform the images from their original size after preprocessing to a suitable input size. First, we follow a same-sized cropping strategy which we employed in the last year's challenge [10]. Here, we take a random crop from the preprocessed image. Second, we follow a random-resize strategy which is popular for ImageNet training [11]. Here, the image is randomly resized and scaled when taking a crop from the preprocessed image.

CNN training

We train all models for 100 epochs using Adam. We use a weighted cross-entropy loss function where underrepresented classes receive a higher weight-based frequency in the training set. Each class is multiplied by a factor $n_i = (N/N_i)^k$ where N is the total number of training images, N_i is the number of images in class i and k controls the balancing severity. We found $k = 1$ to work best. We also tried to use the focal loss [12] with the same balancing weights without performance improvements. Batch size and learning rate are adopted based on GPU memory requirements of each architecture. We halve the learning every 25 epochs. We evaluate every 10 epochs and save the model achieving the best mean sensitivity (best). Also, we save the last model after 100 epochs of training (last). Training is performed on NVIDIA GTX 1080TI (B0-B4) and Titan RTX (B5,B6) graphics cards.

Meta data architecture

For task 2, the meta data is fed into a two-layer neural network with 256 neurons each. Each layer contains batch normalization, a ReLU activation and dropout with $p = 0.4$. The network's output is concatenated with the CNN's feature vector after global average pooling. Then, we apply another layer with batch normalization, ReLU, and dropout. As a baseline we use 1024 neurons which are scaled up for larger models, using EfficientNet's scaling rules for network width. Then, the classification layer follows.

Meta data augmentation

We use a simple data augmentation strategy to address the problem of missing values. During training, we randomly encode each property as missing with a probability of $p = 0.1$. We found this to be necessary as our images for the unknown class do not have any meta data. Thus, we need to ensure that our models do not associate missingness with this class.

Meta data training

During meta data training, the CNN's weights remain fixed. We still employ our CNN data augmentation strategies described above, i.e., the CNN still performs forward passes during training and the CNN's features are not fixed for each image. The meta data layers, i.e., the two-layer network, the layer after concatenation and the classification layer are trained for 50 epochs with a learning rate of 0.00001 and a batch size of 20.

Prediction

After training, we create predictions, depending on the CNN's input strategy. For same-sized cropping, we take 36 ordered crops from the preprocessed image and average the softmaxed predictions of all crops. For random-resize cropping, we perform 16 predictions for each image with four differently scaled center crops and flipped versions of the preprocessed images. For the meta data, we pass the same data through the network for each crop. Again, the softmaxed predictions are averaged.

Ensembling

Finally, we create a large ensemble out of all our trained models. We use a strategy where we select the optimal subset of models based on cross-validation performance [13]. Consider $C = \{c_1, \ldots, c_n\}$ configurations where each configuration uses different hyperparameters (e.g. same-sized cropping) and baseline architectures (e.g. EN B0). Each configuration c_i contains $m = 5$ trained models (best), one for each cross-validation split v_j. We obtain predictions $\hat{y}^i_j$ for each c_i and v_j. Then, we perform an exhaustive search to find $C^* \subseteq C$ such that $\hat{y}^* = \frac{1}{|C^*|} \sum_{i \in C^*} \frac{1}{m} \sum_{j=1}^{m} \hat{y}^i_j$ maximizes the mean sensitivity S.We consider our 8 top performing configurations from the ISIC 2019 Challenge Task 1 in terms of CV performance in C.We perform the search using the best models found during training only but we also include the last models in the final ensemble to have a larger variability. Finally, we obtain predictions for the final test set using all models of all $c_i \in C^*$.

Method validation

For evaluation, we consider the mean sensitivity S for training with images only and for training with additional meta data. The results for cross-validation with individual models and our ensemble are shown in Table 1. Overall, large ENs tend to perform better. Comparing our input strategies, both appear to perform similarly in most cases. Including the ninth class with different skin alterations slightly reduces performance for the first eight classes. Ensembling leads to substantially improved

Table 1
All cross-validation results for different configurations. We consider same-sized cropping (SS) and random-resize cropping (RR) and different model input resolutions. Values are given in percent as mean and standard deviation over all five CV folds. Ensemble average refers to averaging over all predictions from all models. Ensemble optimal refers to averaging over the models we found with our search strategy for the optimal subset of configurations. C = 8 refers to training with eight classes without the unknown class. T1 refers to Task 1 without meta data and T2 refers to Task 2 with meta data. ResNext WSL 1 and 2 refer to ResNeXt-101 WSL 32×8d and 32×16d, respectively [8].

Configuration	Sensitivity T1	Sensitivity T2
SENet154 SS 224×224	67.2 ± 0.8	70.0 ± 0.8
ResNext WSL 1 SS 224×224	65.9 ± 1.6	68.1 ± 1.3
ResNext WSL 2 SS 224×224	65.3 ± 0.8	69.1 ± 1.5
EN B0 SS 224×224 C = 8	66.7 ± 1.8	68.8 ± 1.5
EN B0 SS 224×224	65.8 ± 1.7	67.4 ± 1.6
EN B0 RR 224×224	67.0 ± 1.6	68.9 ± 1.7
EN B1 SS 240×240	65.9 ± 1.6	68.2 ± 1.8
EN B1 RR 240×240	66.8 ± 1.5	68.5 ± 1.8
EN B2 SS 260×260	67.2 ± 1.4	69.0 ± 2.5
EN B2 RR 260×260	67.6 ± 2.0	70.1 ± 2.0
EN B3 SS 300×300	67.8 ± 2.0	68.5 ± 1.7
EN B3 RR 300×300	67.0 ± 1.5	68.4 ± 1.5
EN B4 SS 380×380	67.8 ± 1.1	68.5 ± 1.1
EN B4 RR 380×380	68.1 ± 1.6	69.4 ± 2.2
EN B5 SS 456×456	68.2 ± 0.9	68.7 ± 1.6
EN B5 RR 456×456	68.0 ± 2.2	69.0 ± 1.6
EN B6 SS 528×528	68.8 ± 0.7	69.0 ± 1.4
Ensemble Average	71.7 ± 1.7	73.4 ± 1.6
Ensemble Optimal	72.5 ± 1.7	74.2 ± 1.1
Official Testset	63.6	63.4

performance. Our optimal ensembling strategy improves performance slightly. The optimal ensemble contains nine out of the sixteen configurations.

Regarding meta data, performance tends to improve by 1 to 2% points through the incorporation of meta data. This increase is mostly observed for smaller models as larger models show only minor performance changes. The final ensemble shows improved performance.

For our final submission to the ISIC 2019 Challenge task 1 we created an ensemble with both the best and last model checkpoints. For task 2, we submitted an ensemble with the best model checkpoints only and an ensemble with both best and last model checkpoints. The submission with only the best model checkpoints performed better. Overall, the performance on the official test set is substantially lower than the cross-validation performance. The performance for task2 is lower than the performance for task 1.

Table 2 shows several metrics for the performance on the official test set. For task 1, the performance for the unknown class is substantially lower than for all other classes across several metrics. For task 2, the performance for the unknown class is also substantially reduced, compared to task 1.

Challenge background

Automated skin lesion classification is a challenging problem that is typically addressed using convolutional neural networks. Recently, the ISIC 2018 Skin Lesion Analysis Towards Melanoma Detection challenge resulted in numerous high-performing methods that performed similarly to human experts for the evaluation of dermoscopic images [14]. To improve diagnostic performance further, the ISIC 2019 challenge comes with several old and new problems to consider. In particular, the test set of the ISIC 2019 challenge contains an unknown class that is not present in the dataset.

Table 2
Results from the official test set of the ISIC 2019 Challenge for each class. We consider the AUC, the AUC for a sensitivity larger than 80% (AUC-S), the sensitivity and specificity. Note that the sensitivity given here is differently calculated than S. Values are given in percent.

Class	Task1				Task2			
	AUC	AUC-S	Sens.	Spec.	AUC	AUC-S	Sens.	Spec.
MEL	0.928	0.849	0.594	0.962	0.931	0.849	0.545	0.976
NV	0.96	0.93	0.71	0.975	0.96	0.932	0.637	0.983
BCC	0.949	0.904	0.721	0.94	0.947	0.901	0.649	0.958
AK	0.914	0.824	0.484	0.965	0.919	0.841	0.46	0.966
BKL	0.904	0.805	0.394	0.985	0.908	0.821	0.324	0.991
DF	0.979	0.963	0.578	0.992	0.98	0.965	0.556	0.993
VASC	0.956	0.925	0.644	0.991	0.942	0.912	0.495	0.995
SCC	0.938	0.876	0.439	0.986	0.93	0.878	0.408	0.987
UNK	0.775	0.581	0.00283	0.999	0.612	0.253	0	0.999

Also, the severe class imbalance of real-world datasets is still a major point that needs to be addressed. Furthermore, the training dataset, previously HAM10000 [1], was extended by additional data from the BCN_20000 [2] and MSK dataset [15]. The images have different resolutions and were created using different preprocessing and preparation protocols that need to be taken into account. For challenge task 1, skin lesions have to be classified based on dermoscopic images only. For task 2, dermoscopic images and additional patient meta data have to be used.

Method discussion

We explore multi-resolution EfficientNets for skin lesion classification, combined with extensive data augmentation, loss balancing and ensembling for our participation in the ISIC 2019 Challenge. In previous challenges, data augmentation and ensembling were key factors for high-performing methods [4]. Also, class balancing has been studied [16] where loss weighting with a cross-entropy loss function performed very well. We incorporate this prior knowledge in our approach and also consider the input resolutions as an important parameter. Our results indicate that models with a large input size perform better, see Table 1. For a long time, small input sizes have been popular and the effectiveness of an increased input size is likely tied to EfficientNet's new scaling rules [5]. EfficientNet scales the models' width and depth according to the associated input size which lead to high-performing models with substantially lower computational effort and fewer parameters compared to other methods. We find that these concepts appear to transfer well to the problem of skin lesion classification.

When adding meta data to the model, performance tends to improve slightly for our cross-validation experiments. The improvement is particularly large for smaller, lower-performing models. This might indicate that meta data helps models that do not leverage the full information that is available in the images alone.

The ISIC 2019 Challenge also includes the problem to predict an additional, unknown class. At the point of submission, there was no labeled data available for the class, thus, cross-validation results do not reflect our model's performance with respect to this class. The performance on the official test provides some insights into the unknown class, see Table 2. First, it is clear that the performance on the unknown class is substantially lower than the performance on the other classes. This could explain why there is a substantial difference between our cross-validation results and the results on the official test set. Second, we can observe a substantial performance reduction for the unknown class between task 1 and task 2. This might explain the lack of improvement for task 2, although our cross-validation performance improved with additional meta data. This is likely linked to the fact that we do not have meta data for our unknown class training images. Although we tried to overcome the problem with meta data dropout, our models appear to overfit to the characteristic of missing data for the unknown class.

Overall, we find that EfficientNets perform well for skin lesion classification. In our final ensembling strategy, various EfficientNets were present, although the largest ones performed best. This indicates that a mixture of input resolutions is a good choice to cover multi-scale context for skin lesion classification. Also, SENet154 and the ResNext models were automatically selected for the final ensemble which indicates that some variability in terms of architectures is helpful.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

Publishing fees supported by Funding Programme *Open Access Publishing* of Hamburg University of Technology (TUHH). This work was partially supported by the Forschungszentrum Medizintechnik Hamburg (02fmthh2017).

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.mex.2020.100864.

References

[1] P. Tschandl, C. Rosendahl, H. Kittler, The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions, (eng), Scient. Data 5 (2018) 180161.
[2] M. Combalia et al., Bcn20000: Dermoscopic lesions in the wild, arXiv preprint arXiv:1908.02288, 2019.
[3] J. Kawahara, S. Daneshvar, G. Argenziano, G. Hamarneh, 7-Point Checklist and Skin Lesion Classification using Multi-Task Multi-Modal Neural Nets, (eng), IEEE J. Biomed. Health inf. (2018).
[4] N. Codella et al., Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic), arXiv preprint arXiv:1902.03368, 2019.
[5] M. Tan, Q. Le, EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks, in: Proceedings of the 36th International Conference on Machine Learning, 2019, pp. 6105–6114.
[6] E.D. Cubuk, B. Zoph, D. Mane, V. Vasudevan, Q.V. Le, AutoAugment: Learning augmentation strategies from data, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019.
[7] J. Hu, L. Shen, G. Sun, Squeeze-and-Excitation Networks, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*,, 2018.
[8] D. Mahajan, Exploring the limits of weakly supervised pretraining, in: *Proceedings of the* European Conference on Computer Vision (ECCV), 2018.
[9] T. DeVries and G. W. Taylor, Improved regularization of convolutional neural networks with cutout, arXiv preprint arXiv:1708.04552, 2017.
[10] N. Gessert et al., Skin lesion diagnosis using ensembles, unscaled multi-crop evaluation and loss weighting, arXiv preprint arXiv:1808.01694, 2018.
[11] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. Wojna, Rethinking the Inception Architecture for Computer Vision, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.
[12] T.-Y. Lin, P. Goyal, R. Girshick, K. He, P. Dollar, Focal Loss for Dense Object Detection, in: Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017.
[13] N. Gessert, A. Schlaefer, Left Ventricle Quantification Using Direct Regression with Segmentation Regularization and Ensembles of Pretrained 2D and 3D CNNs, in: M. Pop, et al. (Eds.), Statistical Atlases and Computational Models of the Heart. Multi-Sequence CMR Segmentation, CRT-EPiggy and LV Full Quantification Challenges. STACOM 2019. Lecture Notes in Computer Science, vol 12009, Springer, Cham, 2020, doi:10.1007/978-3-030-39074-7_39.
[14] P. Tschandl, et al., Comparison of the accuracy of human readers versus machine-learning algorithms for pigmented skin lesion classification: An open, web-based, international, diagnostic study, Lancet Oncol. 20 (7) (2019) 938–947.
[15] N.C.F. Codella, et al., Skin lesion analysis toward melanoma detection, in: Proceedings of the International symposium on biomedical imaging (ISBI), hosted by the international skin imaging collaboration (ISIC)," IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018), Washington, DC, 2018, pp. 168–172.
[16] N. Gessert, et al., Skin lesion classification using CNNs with patch-based attention and diagnosis-guided loss weighting, (eng), IEEE Trans. Bio-med. Eng. (2019).