Hindawi
BioMed Research International
Volume 2021, Article ID 6673852, 6 pages
https://doi.org/10.1155/2021/6673852

Research Article

Skin Lesion Classification Using Additional Patient Information

Qilin Sun[1], **Chao Huang**[2], **Minjie Chen**[3], **Hui Xu**[1], **and Yali Yang**[1]

[1] *Department of Dermatology, Shanghai Ninth Hospital affiliated to Shanghai Jiao Tong University, School of Medicine, No. 639, Manufacturing Bureau Road, Huangpu District, Shanghai 200011, China*
[2] *Department of Orthopaedics, West China Hospital of Sichuan University, No. 37 Guoxue Alley, Wuhou District, Chengdu, 610041 Sichuan, China*
[3] *Zeku Technology Co., Ltd., 8th Floor, Building 1, No. 61, Shen Xia Road, Pudong New District Shanghai 201203, China*

Correspondence should be addressed to Hui Xu; xusunan@hotmail.com and Yali Yang; ky7812@126.com

Received 21 October 2020; Revised 28 February 2021; Accepted 19 March 2021; Published 12 April 2021

Academic Editor: Adam Reich

In this paper, we describe our method for skin lesion classification. The goal is to classify skin lesions based on dermoscopic images to several diagnoses' classes presented in the HAM (Human Against Machine) dataset: melanoma (MEL), melanocytic nevus (NV), basal cell carcinoma (BCC), actinic keratosis (AK), benign keratosis (BKL), dermatofibroma (DF), and vascular lesion (VASC). We propose a simplified solution which has a better accuracy than previous methods, but only predicted on a single model that is practical for a real-world scenario. Our results show that using a network with additional metadata as input achieves a better classification performance. This metadata includes both the patient information and the extra information during the data augmentation process. On the international skin imaging collaboration (ISIC) 2018 skin lesion classification challenge test set, our algorithm yields a balanced multiclass accuracy of 88.7% on a single model and 89.5% for the embedding solution, which makes it the currently first ranked algorithm on the live leaderboard. To improve the inference accuracy. Test time augmentation (TTA) is applied. We also demonstrate how Grad-CAM is applied in TTA. Therefore, TTA and Grad-CAM can be integrated in heat map generation, which can be very helpful to assist the clinician for diagnosis.

1. Introduction

Skin cancer is the most common cancer around the world. Early detection and monitoring play a crucial role decreasing the mortality rate of skin cancer. However, it is a challenging problem that only 65%-80% of the skin cancer cases are correctly diagnosed using clinical inspection by an experienced physician [1]. Perez et al. investigated the impact of 13 data augmentation scenarios, such as traditional color and geometric transforms, elastic transforms, random erasing, and lesion mixing method for melanoma classification. The results confirmed that data augmentation can lead to more performance gains than obtaining new images. Recently, a challenge from the international skin imaging collaboration (ISIC) Skin Lesion Analysis Towards Melanoma Detection resulted in numerous high-performing methods that performed similar to human experts for the evaluation of dermoscopic images, which was mostly based on the convolutional neural network technique [2]. And most of those methods obtain a high classification accuracy through ensembling multiple models [3, 4]. For example, 18 convolutional neural network (CNN) architectures [4] and 7 multiresolution EfficientNet (B4) models [3] are explored, with extensive data augmentation. The final results are obtained by 90 submodels, which takes 13.9 seconds to classify single test images for high-end TitanV graphic card [5]. However, in these ensembling method, every image has to be sent through all the models for inference process, and this scheme will even run multiple times on each model for data augmentation, which have a significant amount of computation and would not be practical for a real-world scenario.

Here, we propose a simplified solution which has a better accuracy than previous methods and is also practical for a real-world scenario.

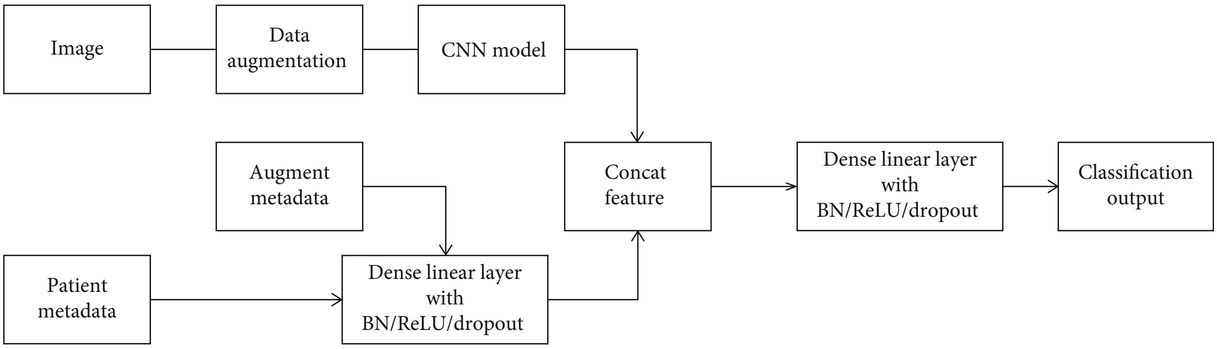


Figure 1: Architecture of the proposed CNN model with metadata.

2. Method Details

2.1. Datasets. The ISIC 2019 skin lesion classification challenge dataset contains 25, 331 dermoscopic images, with extra meta information about the patients' age, the anatomical site, and the sex properties [6–8]. We also use 1,572 extra images in training, including 170 images from the MED-NODE dataset, 533 from the seven-point dataset, 120 from the PH2 dataset, and the remaining ones are our own collected data.

2.2. Image Preprocessing. We resize all images' longer side to 1,024 pixels while preserving the aspect ratio. The shades of gray color constancy method proposed by Finlayson and Trezzi is applied beforehand as a preprocess step, and its color gain in RGB channel is recorded as extra metainfo input. [3]

2.3. Additional Patient Information Preprocessing. The patient's additional patient information is mostly encoded by a one-hot encoding scheme. For example, six features are used for the anatomical site with enough appearance (>100). Sex is encoded as 1/-1/0 for man/female/missing, respectively. Age is encoded by 18 features, under 18 thresholds from 5-90 with step 5, where 1/-1/0 represented larger/smaller/missing, respectively.

2.4. CNN Architectures. We use EfficientNet that have been pretrained on the ImageNet dataset [9]. This model family contains 8 different models that are structurally similar and follow certain scaling rules for adjustment to larger image sizes, from the smallest version B0 to larger versions, up to B7 (Figure 1). To incorporate additional patient information such as age, anatomical site, and sex, an additional dense neural network and fuse its features with the CNN is discussed [4]. In our experiment, we reported the performance of a single model B4, as well as an ensemble model with B3 and B4. We use the default input size described in EfficientNet paper, which is 300×300 for B3 and 380×380 for B4.

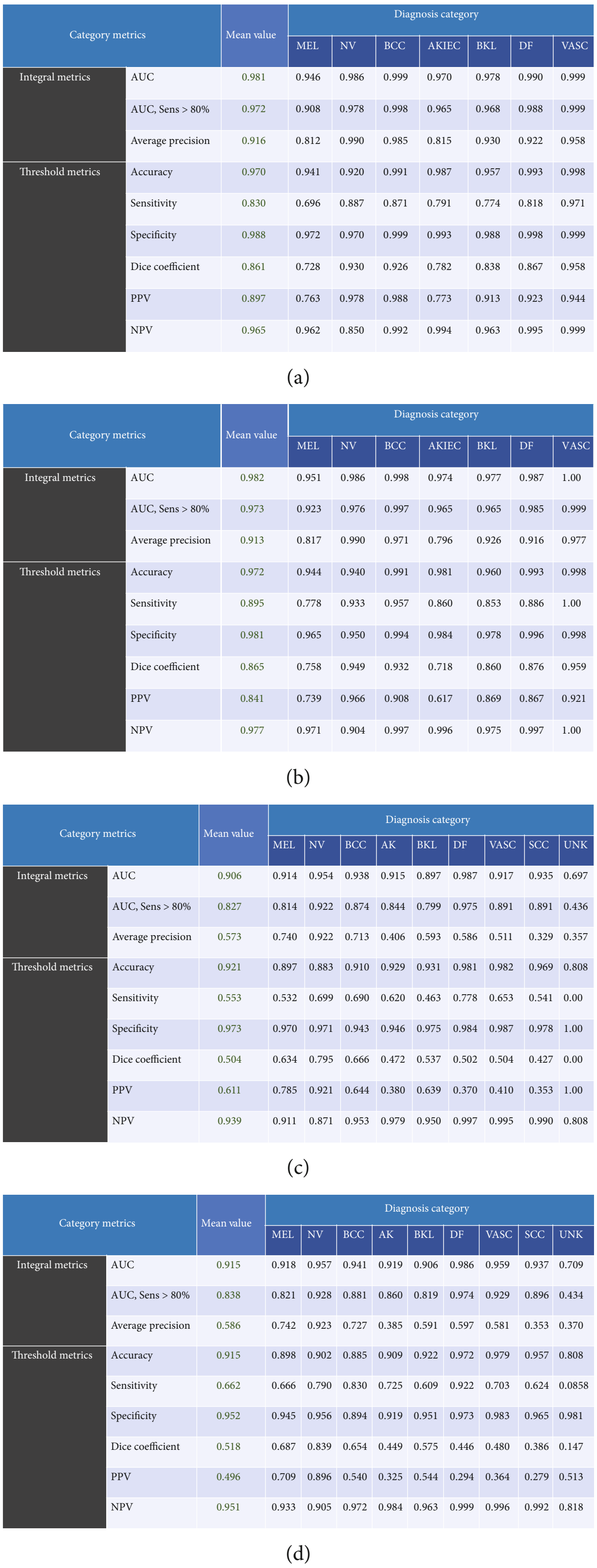
2.5. CNN Data Augmentation. We perform data augmentation in training, both geometric and pixelwise, including random brightness, contrast, hue, saturation, Gaussian Noise, Gaussian blur, random crop, rotation, and flipping. Moreover, we also recorded the scaling and shifting properties of the geometric augmentation which was later used as the metafeatures during the training process. These augmentation processes were implemented for using the albumentations.

2.6. CNN Training. We train the models for 60 epochs with batch size 16 using SGD with momentum. One cycle learning-rate scheduler was applied [10], which was implemented by pytorch OneCycleLR function with default parameter, and the learning rate was set as $1e-3$. A weighted cross-entropy loss function where underrepresented classes receive a higher weight. The coefficient was calculated by the formula described by Lin et al. [11] Focal loss with a gamma = 1.5 was also tested; unfortunately, the testing accuracy was lower [12]. Training was performed on NVIDIA GTX 1080TI. A sampling strategy was applied during the dataloader procedure. In the metafile, a lesion ID was provided for each image. For one lesion, there were 1-30 images in the dataset and those images for the same lesion ID were with high similarity. Therefore, a sampling weight coefficient was added which was equal to inverse of the number of images for that lesion ID.

3. Discussion

Skin cancer is one of the most common malignancy with an increasing incidence rates on a global scale. [13] Early detection is an important factor to increase the overall survival and cure rates for those patients [5]. The diagnosis of those diseases is usually carried out by dermatologists through the visual examination of suspicious skin areas, but it is easy to be misdiagnosed due to the high similarities of some types of lesions. Although supportive imaging techniques such as dermoscopy can improve the accuracy of diagnosis to some extent, the accuracy of diagnosis varies greatly among individuals with different experience.

A large number of studies have been devoted to improving the accuracy of diagnosis and treatment. In recent years, more and more semiautomatic or fully automatic computer-aided diagnosis (CAD) systems based on classical image processing techniques or advanced machine learning paradigms, such as classical workflow of machine learning and CNNs, have been introduced into the diagnosis and treatment of skin diseases as screening procedures or rapid diagnosis tools to assist dermatologists [1, 4, 14]. In addition, the quality of classification could be improved by adding clinical data (such as age, sex, race, skin type and anatomical location) as input to the classifier, and this additional information is helpful for dermatologists to make the right decisions [15]. Perez et al. confirmed that data augmentation can lead to more performance gains than obtaining new images, which was based



Category metrics		Mean value	Diagnosis category						
			MEL	NV	BCC	AKIEC	BKL	DF	VASC
Integral metrics	AUC	0.981	0.946	0.986	0.999	0.970	0.978	0.990	0.999
	AUC, Sens > 80%	0.972	0.908	0.978	0.998	0.965	0.968	0.988	0.999
	Average precision	0.916	0.812	0.990	0.985	0.815	0.930	0.922	0.958
Threshold metrics	Accuracy	0.970	0.941	0.920	0.991	0.987	0.957	0.993	0.998
	Sensitivity	0.830	0.696	0.887	0.871	0.791	0.774	0.818	0.971
	Specificity	0.988	0.972	0.970	0.999	0.993	0.988	0.998	0.999
	Dice coefficient	0.861	0.728	0.930	0.926	0.782	0.838	0.867	0.958
	PPV	0.897	0.763	0.978	0.988	0.773	0.913	0.923	0.944
	NPV	0.965	0.962	0.850	0.992	0.994	0.963	0.995	0.999

(a)

Category metrics		Mean value	Diagnosis category						
			MEL	NV	BCC	AKIEC	BKL	DF	VASC
Integral metrics	AUC	0.982	0.951	0.986	0.998	0.974	0.977	0.987	1.00
	AUC, Sens > 80%	0.973	0.923	0.976	0.997	0.965	0.965	0.985	0.999
	Average precision	0.913	0.817	0.990	0.971	0.796	0.926	0.916	0.977
Threshold metrics	Accuracy	0.972	0.944	0.940	0.991	0.981	0.960	0.993	0.998
	Sensitivity	0.895	0.778	0.933	0.957	0.860	0.853	0.886	1.00
	Specificity	0.981	0.965	0.950	0.994	0.984	0.978	0.996	0.998
	Dice coefficient	0.865	0.758	0.949	0.932	0.718	0.860	0.876	0.959
	PPV	0.841	0.739	0.966	0.908	0.617	0.869	0.867	0.921
	NPV	0.977	0.971	0.904	0.997	0.996	0.975	0.997	1.00

(b)

Category metrics		Mean value	Diagnosis category								
			MEL	NV	BCC	AK	BKL	DF	VASC	SCC	UNK
Integral metrics	AUC	0.906	0.914	0.954	0.938	0.915	0.897	0.987	0.917	0.935	0.697
	AUC, Sens > 80%	0.827	0.814	0.922	0.874	0.844	0.799	0.975	0.891	0.891	0.436
	Average precision	0.573	0.740	0.922	0.713	0.406	0.593	0.586	0.511	0.329	0.357
Threshold metrics	Accuracy	0.921	0.897	0.883	0.910	0.929	0.931	0.981	0.982	0.969	0.808
	Sensitivity	0.553	0.532	0.699	0.690	0.620	0.463	0.778	0.653	0.541	0.00
	Specificity	0.973	0.970	0.971	0.943	0.946	0.975	0.984	0.987	0.978	1.00
	Dice coefficient	0.504	0.634	0.795	0.666	0.472	0.537	0.502	0.504	0.427	0.00
	PPV	0.611	0.785	0.921	0.644	0.380	0.639	0.370	0.410	0.353	1.00
	NPV	0.939	0.911	0.871	0.953	0.979	0.950	0.997	0.995	0.990	0.808

(c)

Category metrics		Mean value	Diagnosis category								
			MEL	NV	BCC	AK	BKL	DF	VASC	SCC	UNK
Integral metrics	AUC	0.915	0.918	0.957	0.941	0.919	0.906	0.986	0.959	0.937	0.709
	AUC, Sens > 80%	0.838	0.821	0.928	0.881	0.860	0.819	0.974	0.929	0.896	0.434
	Average precision	0.586	0.742	0.923	0.727	0.385	0.591	0.597	0.581	0.353	0.370
Threshold metrics	Accuracy	0.915	0.898	0.902	0.885	0.909	0.922	0.972	0.979	0.957	0.808
	Sensitivity	0.662	0.666	0.790	0.830	0.725	0.609	0.922	0.703	0.624	0.0858
	Specificity	0.952	0.945	0.956	0.894	0.919	0.951	0.973	0.983	0.965	0.981
	Dice coefficient	0.518	0.687	0.839	0.654	0.449	0.575	0.446	0.480	0.386	0.147
	PPV	0.496	0.709	0.896	0.540	0.325	0.544	0.294	0.364	0.279	0.513
	NPV	0.951	0.933	0.905	0.972	0.984	0.963	0.999	0.996	0.992	0.818

(d)

Figure 2: The metrics of all lesion type reported on ISIC live leaderboard: (a) our single model on ISIC18, (b) our embedding model on ISIC18, (c) our single model on ISIC19, and (d) our embedding model on ISIC19.

on the researches of the impact of 13 data augmentation scenarios, such as traditional color and geometric transforms, elastic transforms, random erasing, and lesion mixing method for melanoma classification [16]. Although these methods improve the accuracy of diagnosis and treatment, they all have a significant amount of computation and the

Table 1: Results of ISIC 2018 challenge winners from the legacy leaderboard (rows 1-3) and our proposed models (rows 4–6). Among the 16,888 images, there are 15,316 images from ISIC19 dataset, 170 images from the MED-NODE dataset, 533 from the seven-point dataset, 120 from the PH2 dataset, and the remaining data are from our own collected data.

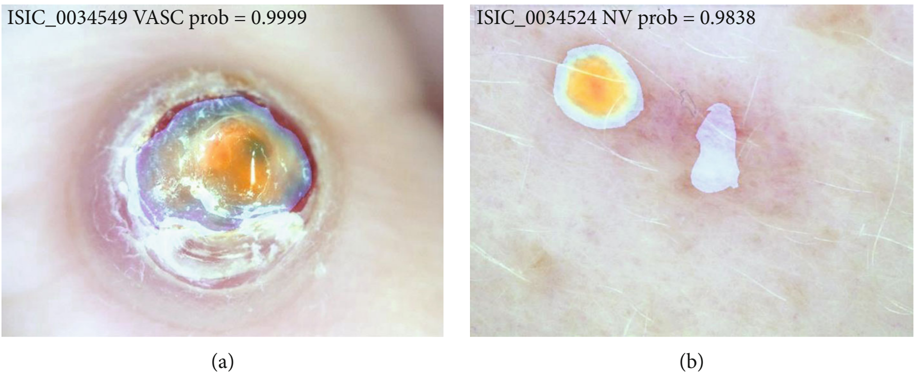
Team/authors	Extra images	BMCA (%)	Sensitivity (%)	Specificity (%)	AUC
Nozdryn et al.	37,807	88.5	83.3	98.6	0.983
Gassert et al. [14]	13,475	85.6	80.9	98.4	0.987
MSM-CNN [5]	2,912	86.2	85.6	97.9	0.987
Our single model (FL)	16,888	88.3	76.1	99.3	0.974
Our single model (CE)	16,888	*88.7*	*83.0*	*98.8*	*0.981*
Our ensemble model	16,888	*89.5*	*89.5*	*98.1*	*0.982*

FL: focal loss; CE: cross-entropy loss.

Table 2: Results of ISIC 2019 challenge winners from the legacy leaderboard (rows 1-3) and our proposed models (rows 4–6). Among the 1,572 images, there are 170 images from the MED-NODE dataset, 533 from the seven-point dataset, 120 from the PH2 dataset, and the remained data are from our own collected data.

Team/authors	Extra images	BMCA (%)	Sensitivity (%)	Specificity (%)	AUC
Gassert et al. [14]	Unknown	63.6	50.7	97.7	0.923
Cancerless	Unknown	63.8	53.1	97.4	0.913
ForCure	Unknown	64.8	53.4	97.4	0.914
Our single model (FL)	1,572	63.9	48.8	97.9	0.899
Our single model (CE)	1,572	*65.0*	*55.3*	*97.3*	*0.906*
Our ensemble model	1,572	*66.2*	*66.2*	*95.2*	*0.915*

FL: focal loss; CE: cross-entropy loss.


ISIC_0034549 VASC prob = 0.9999
ISIC_0034524 NV prob = 0.9838

(a)

(b)

Figure 3: Heat map visualization using Grad-CAM with TTA. The figures are from the ISIC competition (https://challenge2019.isic-archive.com/data.html).

application value in practical operation still needs to be improved.

In this study, we propose a simplified solution and have evaluated our proposed method on both ISIC 2018 and ISIC 2019 test set (Figure 2). The final predicted probability is archived by 10 times TTA, which costs approximated 0.5 second on GTX 1080Ti. There is a live leaderboard to record the performance of the submitted result [2]. The balanced multiclass accuracy (BMCA) is used as the primary metric value, which is shown on Tables 1 and 2. Our results show that using a network with additional patient information as an input achieves a better classification performance. On the ISIC 2018 skin lesion classification challenge test set, our algorithm yields a balanced multiclass accuracy of 88.7% on a single model and 89.5% for the embedding solution, which makes it the currently first ranked algorithm on the live leaderboard and highlights the excellent performance of our proposed solution on this very challenging task (Table 1). Similarly, we also saw the superiority of our proposed method on the ISIC 2019 test set (Table 2). We also observed that weighted cross-entropy criterion achieves higher BMCA than focal loss or label smooth loss criterion.

Grad-CAM is an efficient method to generate the heat map for visualizing where is the hot zone for the classification, which is quite helpful for assisting the clinicians during the diagnosis [17]. To improve the inference accuracy, test time augmentation (TTA) is also applied.

In this paper, we proposed an integrated solution for Grad-CAM and TTA with multicrop. This is implemented by accumulating the heat map on multiple inference process during TTA. As random crop is included in TTA, the final heat map is accumulated for different crop regions with different resolution. As TTA is an efficient way to improve the final prediction metrics; therefore, we believed the weighted heat map generated by this TTA-Grad CAMs operation will also have benefits for the diagnosis of the clinicians.

An example can be seen in Figure 3. The color is the heat map generated by applying Grad-CAM with TTA using the training model. The redder it is, the more likely that the area is a type of disease diagnosed by a neural network. Therefore, TTA and Grad-CAM can be integrated in heat map generation, which can be very helpful to assist the clinician for diagnosis.

We also test the semisupervised scheme as described by utilizing additional unlabeled testing images in the training process. [18] However, we found that the overall accuracy had not been improved. We also tested other advanced augmentation methods, such as Cutmix [19], as well as some attention-based method WS-DAN [20], which achieved the best result in fine-grained image classification tasks, but the performance has not improved.

4. Conclusion

In this study, we have proposed a single baseline for skin lesion classification which uses the information of data augmentation as additional patient information. The metadata used in our manuscript included additional infos that are generated during data augmentation, for example, gain of color normalization process, random crop, and image size properties. Our method has achieved the best result of ISIC live leaderboard with a balanced multiclass accuracy of 88.7% on a single model and 89.5% for the embedding solution, making it the currently first ranked algorithm on the live leaderboard. In addition, it is also practical for real application because of its low computational complexity.

Data Availability

The data used to support the findings of this study are included within the article.

Conflicts of Interest

The authors declare no competing interests.

Authors' Contributions

All authors substantially contributed to the manuscript. Qilin Sun and Chao Huang designed the study, performed the literature review, extracted the data, drew the figures, and organized the tables. Minjie Chen helped to revise the manuscript. Hui Xu and Yali Yang reviewed and edited the manuscript. All authors read and approved the final manuscript. Qilin Sun and Chao Huang contributed equally to this study and share co-first authorship. Hui Xu and Yali Yang share co-corresponding authorship (ORCID: Yali Yang, 0000-0001-8301-9943). Qilin Sun and Chao Huang are co-first authors.

Acknowledgments

This study was supported by the Natural Science Foundation of Shanghai Science and Technology Commission (20ZR1432100), the Industry Support Foundation of Huangpu District, Shanghai, Big Data Construction and Application of Artificial Intelligence "Skin Diagnosis and Treatment" (XK2020007), and the Science Popularization Project of Shanghai Science and Technology Commission (19DZ2305600).

References

[1] T. J. Brinker, A. Hekler, J. S. Utikal et al., "Skin cancer classification using convolutional neural networks: systematic review," *Journal of Medical Internet Research*, vol. 20, no. 10, article e11936, 2018.

[2] P. Tschandl, N. Codella, B. N. Akay 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," *The Lancet Oncology*, vol. 20, no. 7, pp. 938–947, 2019.

[3] G. D. Finlayson and E. Trezzi, "Shades of gray and colour constancy.," in *Color and Imaging Conference*, pp. 37–41, Society for Imaging Science and Technology, 2004.

[4] N. Gessert, M. Nielsen, M. Shaikh, R. Werner, and A. Schlaefer, "Skin lesion classification using ensembles of multi-resolution EfficientNets with meta data," *MethodsX*, vol. 7, 2020.

[5] A. Mahbod, G. Schaefer, C. Wang, G. Dorffner, R. Ecker, and I. Ellinger, "Transfer learning using a multi-scale and multi-network ensemble for skin lesion classification," *Computer Methods and Programs in Biomedicine*, vol. 193, 2020.

[6] N. C. F. Codella, D. Gutman, M. E. Celebi et al., "Skin lesion analysis toward melanoma detection: A challenge at the 2017 International symposium on biomedical imaging (ISBI), hosted by the international skin imaging collaboration (ISIC)," in *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pp. 168–172, Washington, DC, USA, 2018.

[7] M. Combalia, N. C. Codella, V. Rotemberg et al., "BCN20000: Dermoscopic lesions in the wild," 2019, http://arxiv.org/abs/1908.02288.

[8] P. Tschandl, C. Rosendahl, and H. Kittler, "The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions," *Scientific Data*, vol. 5, no. 1, 2018.

[9] M. Tan and Q. V. Le, "EfficientNet: rethinking model scaling for convolutional neural networks," in *Proceedings of the 36th International Conference on Machine Learning*, pp. 6105–6114, Long Beach, California, USA, 2019.

[10] L. N. Smith and N. Topin, "Super-convergence: very fast training of neural networks using large learning rates," in *Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications*, Baltimore, Maryland, United States, 2019.

[11] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal loss for dense object detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 2, pp. 318–327, 2020.

[12] Y. Cui, M. Jia, T.-Y. Lin, Y. Song, and S. Belongie, "Class-balanced loss based on effective number of samples," in *2019*

IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 9260–9269, Long Beach, CA, USA, 2019.

[13] U. Leiter, T. Eigentler, and C. Garbe, "Epidemiology of Skin Cancer," in *Sunlight, Vitamin D and Skin Cancer*, pp. 120–140, Springer, New York, NY, USA, 2014.

[14] N. Gessert, T. Sentker, F. Madesta et al., "Skin lesion diagnosis using ensembles, unscaled multi-crop evaluation and loss weighting," 2018, http://arxiv.org/abs/1808.01694.

[15] H. Haenssle, C. Fink, R. Schneiderbauer et al., "Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists," *Annals of oncology*, vol. 29, no. 8, pp. 1836–1842, 2018.

[16] F. Perez, C. Vasconcelos, S. Avila, and E. Valle, "Data uugmentation for skin lesion analysis," in *International Workshop on Computer-Assisted and Robotic Endoscopy Workshop on Clinical Image-Based Procedures International Workshop on OR 2.0 Context-Aware Operating Theaters International Workshop on Skin Image Analysis*, pp. 303–311, Springer, 2018.

[17] P. Tschandl, C. Rinner, Z. Apalla et al., "*Human-computer collaboration for skin cancer recognition*," *Nature Medicine*, vol. 26, no. 8, pp. 1229–1234, 2020.

[18] K. Sohn, D. Berthelot, C. L. Li et al., "FixMatch: simplifying semi-supervised learning with consistency and confidence," 2020, http://arxiv.org/abs/2001.07685.

[19] S. Yun, D. Han, S. J. Oh, S. Chun, J. Choe, and Y. Yoo, "CutMix: regularization strategy to train strong classifiers with localizable features," in *International Conference on Computer Vision*, pp. 6023–6032, Seoul, South Korea, 2019.

[20] T. Hu, H. Qi, Q. Huang, and Y. Lu, "See better before looking closer: weakly supervised data augmentation network for fine-grained visual classification," 2019, http://arxiv.org/abs/1901.09891.