Bulk Production Augmentation Towards Explainable Melanoma Diagnosis

Kasumi Obi
Applied Informatics, Graduate School of
Science and Engineering
Hosei University, Tokyo, Japan
Email: kasumi.obi.6e@stu.hosei.ac.jp

Quan Huu Cap
Applied Informatics, Graduate School of
Science and Engineering
Hosei University, Tokyo, Japan
Email: huu.quan.cap.78@stu.hosei.ac.jp

Noriko Umegaki-Arao
Department of Dermatology,
Tokyo Women's Medical University
Medical Center East, Tokyo, Japan
Email: umegaki.noriko@twmu.ac.jp

Masaru Tanaka
Department of Dermatology,
Tokyo Women's Medical University
Medical Center East, Tokyo, Japan
Email: masarutanaka@1984.jukuin.keio.ac.jp

Hitoshi Iyatomi
Applied Informatics, Graduate School of
Science and Engineering, Hosei University, Tokyo, Japan
Email: iyatomi@hosei.ac.jp

Abstract—Although highly accurate automated diagnostic techniques for melanoma have been reported, the realization of a system capable of providing diagnostic evidence based on medical indices remains an open issue because of difficulties in obtaining reliable training data. In this paper, we propose bulk production augmentation (BPA) to generate high-quality, diverse pseudo-skin tumor images with the desired structural malignant features for additional training images from a limited number of labeled images. The proposed BPA acts as an effective data augmentation in constructing the feature detector for the atypical pigment network (APN), which is a key structure in melanoma diagnosis. Experiments show that training with images generated by our BPA largely boosts the APN detection performance by 20.0 percentage points in the area under the receiver operating characteristic curve, which is 11.5 to 13.7 points higher than that of conventional CycleGAN-based augmentations in AUC.

Index Terms—data augmentation, 7-point checklist, diagnostic system, generative adversarial networks.

I. INTRODUCTION

Melanoma is the most serious type of skin cancer, and it progresses very quickly. The WHO reported 287,723 new cases and 60,712 fatalities in 2018 [1]. Melanoma is normally diagnosed using dermoscopy (a dermatological microscope). However, it requires dermatologist experience, and this sometimes lacks objectivity [2]. For these reasons, many studies on the development of autonomous melanoma diagnosis systems using dermoscopy images have been proposed.

Prior to the advancement of deep learning techniques, a typical system using classical pattern recognition consists of the following three steps: segmentation of tumor areas [3], [4], feature extraction based on medical knowledge, and the final classification [5]–[7]. There have also been some attempts to estimate well-known medical indicators [8]–[11], such as the ABCD rule [12] and the 7-point checklist [13].

Despite achieving excellent results, most of these techniques are based on small datasets, so their practicality and generalization performance are still questionable. Recently,

many studies have reported excellent results, thanks to the rapid progress of deep learning techniques and the release of large-scale dermoscopy image datasets [14]–[16]. For instance, Gessert et al. proposed a high-accuracy skin lesion diagnosis system [17] that advances ensemble learning of multiple deep learning models. Kitada et al. proposed an unlabeled dermoscopy images classifier [18] using semi-supervised learning and body hair augmentation.

Providing evidence for diagnosis results is very crucial for making medical decisions, such as the diagnosis of cancer. Model interpretability has become a more important topic in recent deep learning studies. Several studies have proposed explainable melanoma diagnosis systems by utilizing Grad-CAM [19] to provide potential malignant regions [20] or by improving the intuitiveness of heatmaps to make the model more explainable [21]. However, the resulting heatmap is not necessarily interpretable for users because it is a display based on the gradient of the discriminator obtained by training and is not based on medical knowledge. We believe that providing quantitative diagnosis evidence, such as the 7-point checklist, can improve model interpretability.

As a related study, Murabayashi et al. [22] proposed a method to quantitatively estimate each measure of the 7-point checklist by taking advantage of semi-supervised learning with virtual adversarial training [23] and multi-task learning [24]. They trained their model on only 226 labeled images with the 7-point checklist by four dermatologists and 9,124 images with labels of benign/malignant information. As a result, they were able to estimate each criterion on the 7-point checklist with a deviation equivalent to that of a dermatologist, and the diagnostic accuracy was significantly improved by 7.5%. However, the quantitative and extensive assessment of these clinical indicators is still limited because of the lack of large-scale images labeled with these indicators [22], [25]. Labeling large datasets with multiple dermatologists is expensive, and obtaining such datasets is difficult. Thus, addressing the lack

of dataset could improve model performance.

Recently, models based on generative adversarial networks (GANs) [26] have been successfully used in various fields. In the application for computer vision, progressive growing GAN (PGGAN) [27] achieved the synthesis of high-resolution realistic images by progressively enlarging image resolution during the training process. CycleGAN [28] learns the relationship between images from two different domains to translate styles without paired training images; this process has been applied in a variety of fields.

Some studies have reported improving medical diagnosis accuracy by increasing the number of training images using GAN-based image generation models with a limited number of training data [29], [30]. However, since GAN learns to estimate the distribution of training data, it cannot expand the distribution itself because the generated image is basically an interpolated sample from the dataset [31]. This also applies to CycleGAN; the variety of generated images is limited because CycleGAN can only perform one-to-one conversion.

Therefore, in this paper, we propose bulk production augmentation (BPA), which is a framework for generating a large volume and variety of pseudo-images of the desired category, and we apply it to the quantification task of the dermoscopic structure toward an explainable melanoma diagnosis. The proposed BPA can be divided into two phases: (1) bulk production phase and (2) feature transition phase. In the bulk production phase, PGGAN augments high-resolution images of the nevus, which are widely available and have less malignant features. They are used as the foundation for adding malignant features in the second phase. Then, in the feature transition phase, malignant features are added to the generated nevus images using CycleGAN. With these steps, the system can generate a large number of varied dermoscopy images with malignant features.

In the evaluation experiments, we tested the effectiveness of the proposed BPA framework on the atypical pigment network (APN), an indicator registered as a major criterion in the 7-point checklist. For the training images, we used 230 images labeled as APN and 10,000 nevus images from public datasets.

With respect to the evaluation of our proposed data generation (i.e., augmentation) framework, direct evaluation by dermatologists is extremely expensive and infeasible. We therefore trained two types of classifiers to confirm if we could improve classification performance by augmenting the data with our framework. The first type of classifier is the detectors of APN structures, and we used them to verify how data augmentation by BPA improves detection performance. The second type is a melanoma-nevus classifier, which is used to verify whether the images generated by the APN's constructional features by BPA improves the grade of malignancy compared with that before the assignment.

II. METHOD

A. The 7-point checklist - preparation

The 7-point checklist is a well-known diagnostic method using dermoscopy. This checklist requires the identification of

TABLE I
7-POINT CHECKLIST CRITERIA

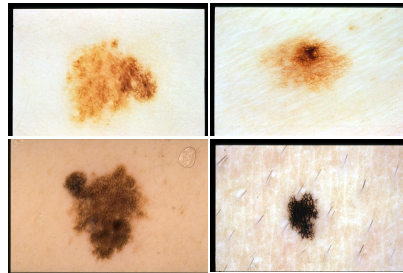| Major criteria	weight
1. Atypical pigment network	×2
2. Blue-whitish veil	×2
3. Atypical vascular pattern	×2
Minor criteria	
4. Irregular streaks	×1
5. Irregular pigmentation	×1
6. Irregular dot/globules	×1
7. Regression structures	×1

Fig. 1. Dermoscopy images with atypical pigment network (APN) structures.

seven dermoscopic structures, as shown in Table I. The first three are categorized as major criteria, and the latter four are minor criteria. The score for a skin lesion is determined as the weighted sum of the structures present in it. Using the 7-point checklist, the total score (TS) is calculated as

$$TS = \#major \times 2 + \#minor, \tag{1}$$

where #major and #minor are the number of major and minor dermoscopic structures present in the image, respectively. Accordingly, the TS ranges between 0 and 10. If the TS is greater than or equal to 3, the lesion is considered malignant. According to the literature, diagnostic accuracy based on the 7-point checklist by 40 experts is 75.0% in sensitivity and 76.2% in specificity [2]. In this paper, we focus on the classification of the APN as one of the major criteria. Fig. 1 shows a typical example of dermoscopy images with APN structures. While the APN is a strong indicator of melanoma diagnosis, it should be noted that the presence of this structure does not necessarily indicate malignancy.

B. Bulk production augmentation（BPA）

In this paper, we propose the bulk production augmentation framework (BPA) in order to develop an interpretable automated melanoma diagnosis system. The proposed BPA is a novel and effective data augmentation framework when there are very little supervised data for training. Our BPA is composed of two phases: (1) the bulk production phase and (2) the feature transition phase. Fig. 2 shows the flow of our proposed BPA framework. Each phase uses different kinds
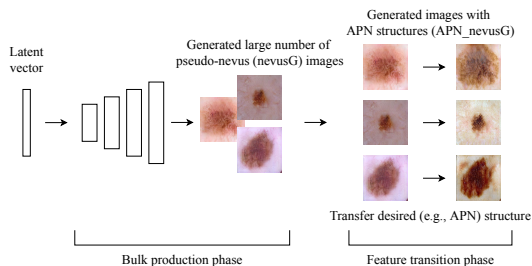
Fig. 2. The overview of the proposed bulk production augmentation (BPA).

of GANs. The first phase involves the generation of a large number of nevus images, and the second phase involves the transfer of APN features to the diverse nevus images generated in the first phase.

1) The bulk production phase: The bulk production phase generates a number of diverse images that serve as the basis for the characterization in the second phase. In this paper, a nevus (i.e., a benign pigmented skin tumor) is used as the base image. A nevus has a low proportion of malignant dermoscopic features and is readily available in large quantities. It also has attractive attributes as a basis for transferring disease features. In this phase, based on the nevus images, a large number of pseudo-nevus images are generated using an adversarial generation network to further increase diversity. We used PGGAN [27] for image generation. PGGAN takes a latent vector (noise) as an input and gradually scales up the image size, eventually generating various high-quality images.

2) The feature transition phase: The feature transition phase assigns the desired features to the images generated in the previous bulk production phase. Specifically, in this paper, CycleGAN [28] is used to assign APN features to the nevus training datasets (see Table II). CycleGAN learns the data distribution of two domains to achieve a style transformation without the need for image pairs.

The proposed BPA solves the problem of quantity and diversity of training images. The bulk generation stage is the cornerstone of this proposal's data augmentation, as it allows for the generation of virtually unlimited base images.

III. EXPERIMENTS

In this paper, we used the ISIC 2019 open dermoscopy image dataset [14]–[16] and an independent dataset presented in the literature [32]. Specifically, the ISIC 2019 dataset consists of 12,875 nevi and 4,522 melanomas. The latter dataset contains 230 images that were identified as having APN features. For our experiments, the resolution of images to be used in both phases of the BPA was set to 256×256 pixels.

A. Training the BPA

For the training of PGGAN in the bulk production phase, a total of 6,816 images of nevi without artifacts, such as hairs, measures, and pen writings, were selected from the ISIC 2019 dataset to generate high-quality skin tumor images (i.e., pseudo-nevus images).

For the training data of CycleGAN in the feature transition phase, we used the aforementioned 230 dermoscopy images with APN features and 2,000 nevi (randomly collected) from the ISIC 2019 dataset, considering the balance in the number of data.

We followed the default settings from both PGGAN and CycleGAN to train our models. Please refer to the original literature for more training details [27], [28].

B. Evaluation the BPA

The diagnosis of APN features is known to vary widely even among skilled dermatologists, as mentioned in [2]. Therefore, the of the generation of dermoscopy images requires agreement on the diagnosis of multiple dermatologists. However, this is impractical for a large amount of output results. We thus decided to perform the following two indirect but quantitative evaluations (1) Effect on APN detection; (2) Effect on malignancy grading.

1) Effect on APN detection: To evaluate the effect of the proposed BPA framework, deep neural network classifiers for estimating the presence or absence of APN features were trained on several groups of training data, and their performance was compared. As for the evaluation dataset, 134 and 100 dermoscopy images with and without APNs, respectively, were selected from the ISIC 2019 dataset based on the agreement of two skilled dermatologists. Note here that we excluded pigmented skin lesions in acral volar regions because they have completely different dermoscopic structures [33], [34]. The APN classifiers were fine-tuned based on the EfficientNet-B1 network [35], which has been reported to have excellent classification performance. A comparison of the conditions of the training data, including the proposed BPA, is shown in Table II.

For the sake of brevity, the following dataset notation will be used thereafter:
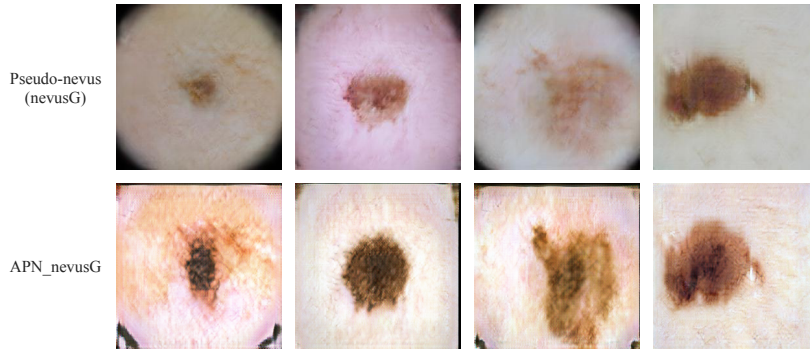
Fig. 3. Example of generated dermoscopy images. The upper images are generated nevus images by the bulk production phase (nevusG). The lower images are transformed with the APN structure from the images by the feature transition phase (APN_nevusG).

- **nevus**: Real nevus images (original training data)
- **nevusG**: Generated pseudo-nevus images from the bulk production phase (used as base skin tumor images in the proposed BPA)
- **APN**: Real images identified with APN structures (original training data)
- **APN_nevus**: Generated images with APN structures by CycleGAN from real nevus images
- **APN_nevusG**: Generated images with APN structures by CycleGAN from generated nevus images (nevusG)

Note here that "APN_nevusG" is the main outcome of our proposed BPA.

The conditions (i.e., training datasets for the APN classifiers) can be summarized as follows:

(A) **baseline**: We used 10,000 real nevus images and 230 real APN images.
(B) **CycleGAN**: In addition to the baseline, 10,000 generated APN images from nevus images (APN_nevus) are added to the training.
(C) **simplified BPA**: In addition to the baseline, 10,000 generated APN images from nevusG (APN_nevusG) are added to the training.
(D) **(proposed) BPA**: In addition to the baseline, 10,000 generated nevus images (nevusG) and 20,000 APN images (APN_nevusG) are added to the training.

The condition (C) is a simplified version of BPA that we introduced to compare the effect of data augmentation by BPA with that of the conventional CycleGAN method. With this condition, we focused on evaluating the efficiency of generating diverse images by the bulk production phase.

2) Effect on malignancy grading: We evaluated the validity of the generated APN feature-assigned images from a different perspective. Because APN features are an indicator of a melanoma diagnosis, an improvement in malignancy should be expected if APN features are properly assigned by BPA. In this experiment, we built a deep classifier capable of classifying either melanoma or nevus with high accuracy, and we evaluated whether the classifier increases the detection degree of malignancy by assigning APN features. To this end, we used the EfficientNet-B1 [35] network as the backbone of

our grading classifier. The model was trained on 8,000 nevi and 4,000 melanomas from the ISIC dataset, and we confirmed that it achieved a sensitivity of 88.8% and a specificity of 94.2% (F1-score: 0.914) for 500 test cases each of nevus and melanoma.

Our classifier used the sigmoid function as its output, so the malignancy grading score ranges from 0.0 to 1.0 in our experiment. We then compared the grading performance of the trained classifier on the following five datasets: real nevus, generated nevus (nevusG), generated APN images from the real nevus (APN_nevus), generated APN images from the generated nevus (APN_nevusG), and real APN images.

3) Training details of the classifiers: All the above classifiers for evaluating our BPA were fined-tuned from the pre-trained EfficientNet-B1 model. The input images were resized to 240×240 pixels, which was optimally designed for EfficientNet-B1. Random resize cropping and random horizontal flip data augmentation were performed with a scale in the range of 0.5 to 1.0 during the training. Furthermore, RandAugment [36], a state-of-the-art data augmentation method, was applied to all models with $N = 6$ and $M = 8$ based on the results of preliminary experiments. Since the datasets in Table II are extremely imbalanced, we introduced the weighted loss [37] to address the class-imbalanced problem. The momentumSGD [38] optimizer with a learning rate of 1×10^{-5} and weight decay for each epoch of 1×10^{-6} are used to train all classifiers. For more details of the configurations and training, please refer to the original EfficientNet paper [35].

IV. RESULTS & DISCUSSION

A. Improving APN feature detection with BPA

Fig. 3 shows the example images generated by the proposed BPA. The generated images are visually natural. A summary of the detection performance for APN features is shown in Table III, and their receiver operating characteristic (ROC) curves are shown in Fig. 4.

For the baseline (A), with a small number of real APN images, the state-of-the-art EfficientNet is unable to detect the APN structure despite the introduction of RandAugment data

TABLE III
COMPARISON OF THE DETECTION PERFORMANCE FOR ATYPICAL PIGMENT NETWORK (APN) STRUCTURES

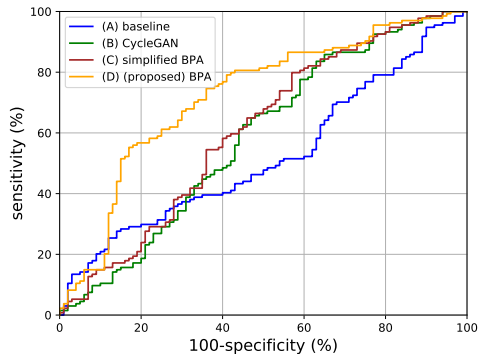| Dataset	Accuracy (%)	Recall (%)	Precision (%)	F1 (%)	AUC
(A) baseline	53.8	69.4	58.1	63.3	0.522
(B) CycleGAN	59.8	64.9	64.9	64.9	0.585
(C) simplified BPA	58.5	54.5	70.0	60.1	0.607
(D) (proposed) BPA	**69.7**	**77.6**	**71.7**	**74.6**	**0.722**

Fig. 4. Receiver operating characteristic curve for detecting the APN structure. Our proposed bulk production augmentation (BPA) largely improved detection performance.
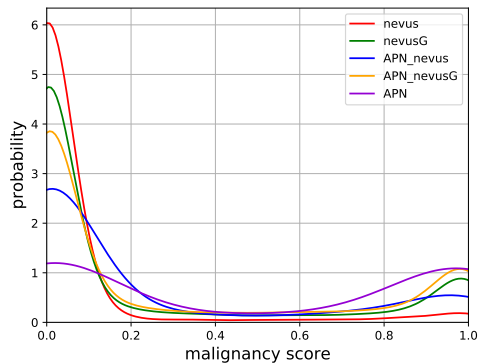

Fig. 5. Comparison of malignancy score distributions. The images obtained by the proposed BPA show a malignancy distribution more similar to that of the APN.

augmentation and the weighted loss. Two CycleGAN-based models (B and C) improve the APN detection performance by 6.3 and 8.5 points in the area under the ROC curve (AUC), respectively, whereas the proposed BPA (D) shows a significant improvement by 20.0 points (0.722 in the AUC). With the introduction of CycleGAN (B), there is an improvement in accuracy, but the accuracy obtained is insufficient. The result from (C), the simplified version of the proposed BPA, which assigns APN features to the pseudo-nevus images (nevusG), shows a slight improvement from (B). This may be due to the generated images being different from any of the original images, making them more effective in the data augmentation process. In the proposed framework (D), BPA successfully generates a large number of diverse and high-quality images with APN structures (APN_nevusG) based on a large number of *base* pseudo-nevus images (nevusG) generated in the bulk production phase. These BPA features are the reason for the significant improvement in APN detection.

B. Assessing the impact of APN features on grading malignancy

Fig. 5 shows the distribution of malignancy in each image by the melanoma-nevus discriminator. A value close to 1 indicates a higher grade of malignancy. While the result on the real APNs has score peaks at 0 and 1, the other results only have large peaks at 0. This is because a certain number of nevi have APN features.

The score distribution of the generated APN_nevusG images by our BPA (yellow) shows higher malignancy than that of the real nevus images (red) and nevusG images (green). It (yellow) shows a higher probability than APN_nevus (blue) in both

areas that are close to grades 0 (benign) and 1 (melanoma), confirming that the APN features are appropriately assigned.

Overall, the proposed BPA can assign malignancy characteristics at about the same level as when the feature translation is performed from real data. The validity of the proposed framework is also confirmed by the fact that the malignancy distribution of the proposal (yellow) is closest to the distribution with real APN features (purple). Here, we need to emphasize that our proposed framework differs from conventional methods because it can generate a large number of images with the desired features, which can be a useful data augmentation method. This advantage led to the significant improvement in the detection of APN features in our experiments. This supports the fact that the above (B) is less effective than our proposed (C) and (D).

V. CONCLUSION

In this paper, we proposed the bulk production augmentation (BPA) for the realization of an automated diagnosis system that can present diagnostic evidence based on medical indices. BPA first produces a large number of data with general features to be generated and then transfers the desired features from the previous diverse generated base images. BPA is a novel and effective data augmentation framework for improving the interpretability of automated medical diagnosis systems when there are very little supervised data for training. We believe that our BPA can be applied for augmenting other dermoscopic structures data defined in the 7-point checklist or other criteria.

References

[1] WHO, "Global cancer observatory," *World Health Organization*. [Online]. Available: https://gco.iarc.fr/

[2] G. Argenziano, H. P. Soyer, S. Chimenti, R. Talamini, R. Corona, F. Sera, M. Binder, L. Cerroni, G. De Rosa, G. Ferrara *et al.*, "Dermoscopy of pigmented skin lesions: results of a consensus meeting via the internet," *Journal of the American Academy of Dermatology*, vol. 48, no. 5, pp. 679–693, 2003.

[3] H. Iyatomi, H. Oka, M. Saito, A. Miyake, M. Kimoto, J. Yamagami, S. Kobayashi, A. Tanikawa, M. Hagiwara, K. Ogawa *et al.*, "Quantitative assessment of tumour extraction from dermoscopy images and evaluation of computer-based extraction methods for an automatic melanoma diagnostic system," *Melanoma Research*, vol. 16, no. 2, pp. 183–190, 2006.

[4] M. E. Celebi, H. Iyatomi, G. Schaefer, and W. V. Stoecker, "Lesion border detection in dermoscopy images," *Computerized Medical Imaging and Graphics*, vol. 33, no. 2, pp. 148–153, 2009.

[5] K. Hoffmann, T. Gambichler, A. Rick, M. Kreutz, M. Anschuetz, T. Grünendick, A. Orlikov, S. Gehlen, R. Perotti, L. Andreassi *et al.*, "Diagnostic and neural analysis of skin cancer (danaos). a multicentre study for collection and computer-aided analysis of data from pigmented skin lesions using digital dermoscopy," *British Journal of Dermatology*, vol. 149, no. 4, pp. 801–809, 2003.

[6] H. Iyatomi, H. Oka, M. E. Celebi, M. Hashimoto, M. Hagiwara, M. Tanaka, and K. Ogawa, "An improved internet-based melanoma screening system with dermatologist-like tumor area extraction algorithm," *Computerized Medical Imaging and Graphics*, vol. 32, no. 7, pp. 566–579, 2008.

[7] H. Ganster, P. Pinz, R. Rohrer, E. Wildling, M. Binder, and H. Kittler, "Automated melanoma recognition," *IEEE Transactions on Medical Imaging*, vol. 20, no. 3, pp. 233–239, 2001.

[8] W. Barhoumi and A. Baâzaoui, "Pigment network detection in dermatoscopic images for melanoma diagnosis," *IRBM*, vol. 35, no. 3, pp. 128–138, 2014.

[9] K. Thon, H. Rue, S. O. Skrøvseth, and F. Godtliebsen, "Bayesian multiscale analysis of images modeled as gaussian markov random fields," *Computational Statistics & Data Analysis*, vol. 56, no. 1, pp. 49–61, 2012.

[10] M. E. Celebi, H. Iyatomi, W. V. Stoecker, R. H. Moss, H. S. Rabinovitz, G. Argenziano, and H. P. Soyer, "Automatic detection of blue-white veil and related structures in dermoscopy images," *Computerized Medical Imaging and Graphics*, vol. 32, no. 8, pp. 670–677, 2008.

[11] M. Sadeghi, T. K. Lee, D. McLean, H. Lui, and M. S. Atkins, "Detection and analysis of irregular streaks in dermoscopic images of skin lesions," *IEEE Transactions on Medical Imaging*, vol. 32, no. 5, pp. 849–861, 2013.

[12] W. Stolz, "Abcd rule of dermatoscopy: a new practical method for early recognition of malignant melanoma," *European Journal of Dermatology*, vol. 4, pp. 521–527, 1994.

[13] G. Argenziano, G. Fabbrocini, P. Carli, V. De Giorgi, E. Sammarco, and M. 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*, vol. 134, no. 12, pp. 1563–1570, 1998.

[14] M. Combalia, N. C. Codella, V. Rotemberg, B. Helba, V. Vilaplana, O. Reiter, C. Carrera, A. Barreiro, A. C. Halpern, S. Puig *et al.*, "Bcn20000: Dermoscopic lesions in the wild," *CoRR, abs/1908.02288*, 2019.

[15] 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, p. 180161, 2018.

[16] N. C. Codella, D. Gutman, M. E. Celebi, B. Helba, M. A. Marchetti, S. W. Dusza, A. Kalloo, K. Liopyris, N. Mishra, H. Kittler *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)," *IEEE International Symposium on Biomedical Imaging*, pp. 168–172, 2018.

[17] N. Gessert, M. Nielsen, M. Shaikh, R. Werner, and A. Schlaefer, "Skin lesion classification using ensembles of multi-resolution efficientnets with meta data," *MethodsX*, p. 100864, 2020.

[18] S. Kitada and H. Iyatomi, "Skin lesion classification with ensemble of squeeze-and-excitation networks and semi-supervised learning," *CoRR, abs/1809.02568*, 2018.

[19] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," *IEEE International Conference on Computer Vision*, pp. 618–626, 2017.

[20] S. S. Han, M. S. Kim, W. Lim, G. H. Park, I. Park, and S. E. Chang, "Classification of the clinical images for benign and malignant cutaneous tumors using a deep learning algorithm," *Journal of Investigative Dermatology*, vol. 138, no. 7, pp. 1529–1538, 2018.

[21] J. Zhang, Y. Xie, Y. Xia, and C. Shen, "Attention residual learning for skin lesion classification," *IEEE Transactions on Medical Imaging*, vol. 38, no. 9, pp. 2092–2103, 2019.

[22] S. Murabayashi and H. Iyatomi, "Towards explainable melanoma diagnosis: Prediction of clinical indicators using semi-supervised and multi-task learning," *IEEE International Conference on Big Data*, pp. 4853–4857, 2019.

[23] T. Miyato, S. I. Maeda, M. Koyama, and S. Ishii, "Virtual adversarial training: a regularization method for supervised and semi-supervised learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 8, pp. 1979–1993, 2018.

[24] R. Caruana, "Multitask learning," *Machine Learning*, vol. 28, no. 1, pp. 41–75, 1997.

[25] H. Iyatomi, H. Oka, M. E. Celebi, M. Tanaka, and K. Ogawa, "Parameterization of dermoscopic findings for the internet-based melanoma screening system," *IEEE Symposium on Computational Intelligence in Image and Signal Processing*, pp. 189–193, 2007.

[26] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *Advances in Neural Information Processing Systems*, pp. 2672–2680, 2014.

[27] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of gans for improved quality, stability, and variation," *International Conference on Learning Representations*, pp. 1–26, 2017.

[28] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," *IEEE International Conference on Computer Vision*, pp. 2223–2232, 2017.

[29] M. Frid-Adar, I. Diamant, E. Klang, M. Amitai, J. Goldberger, and H. Greenspan, "Gan-based synthetic medical image augmentation for increased cnn performance in liver lesion classification," *Neurocomputing*, vol. 321, pp. 321–331, 2018.

[30] C. Han, L. Rundo, R. Araki, Y. Furukawa, G. Mauri, H. Nakayama, and H. Hayashi, "Infinite brain mr images: Pggan-based data augmentation for tumor detection," *Neural Approaches to Dynamics of Signal Exchanges*, pp. 291–303, 2020.

[31] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training gans," *Advances in Neural Information Processing Systems*, pp. 2234–2242, 2016.

[32] J. Kawahara, S. Daneshvar, G. Argenziano, and G. Hamarneh, "Seven-point checklist and skin lesion classification using multitask multimodal neural nets," *IEEE Journal of Biomedical and Health Informatics*, vol. 23, no. 2, pp. 538–546, 2018.

[33] T. Saida, A. Miyazaki, S. Oguchi, Y. Ishihara, Y. Yamazaki, S. Murase, S. Yoshikawa, T. Tsuchida, Y. Kawabata, and K. Tamaki, "Significance of dermoscopic patterns in detecting malignant melanoma on acral volar skin: results of a multicenter study in japan," *Archives of Dermatology*, vol. 140, no. 10, pp. 1233–1238, 2004.

[34] H. Iyatomi, H. Oka, M. E. Celebi, K. Ogawa, G. Argenziano, H. P. Soyer, H. Koga, T. Saida, K. Ohara, and M. Tanaka, "Computer-based classification of dermoscopy images of melanocytic lesions on acral volar skin," *Journal of Investigative Dermatology*, vol. 128, no. 8, pp. 2049–2054, 2008.

[35] M. Tan and Q. V. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," *International Conference on Machine Learning*, pp. 6105–6114, 2019.

[36] E. D. Cubuk, B. Zoph, J. Shlens, and Q. V. Le, "Randaugment: Practical automated data augmentation with a reduced search space," *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 702–703, 2020.

[37] C. Huang, Y. Li, C. C. Loy, and X. Tang, "Learning deep representation for imbalanced classification," *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5375–5384, 2016.

[38] N. Qian, "On the momentum term in gradient descent learning algorithms," *Neural Networks*, vol. 12, no. 1, pp. 145–151, 1999.