Assessing GAN-Based Generative Modeling on Skin Lesions Images

Sandra Carrasco Limeros[1,2], Sylwia Majchrowska[1,2(✉)], Mohamad Khir Zoubi[1], Anna Rosén[1], Juulia Suvilehto[1], Lisa Sjöblom[1], and Magnus Kjellberg[1]

[1] Sahlgrenska University Hospital, Blå stråket 5, 413 45 Göteborg, Sweden
[2] AI Sweden, Lindholmspiren 3-5, 402 78 Göteborg, Sweden
{sandra.carrasco,sylwia.majchrowska}@ai.se

Abstract. We explored unconditional and conditional Generative Adversarial Networks (GANs) in centralized and decentralized settings. The centralized setting imitates studies on large but highly unbalanced skin lesion dataset, while the decentralized one simulates a more realistic hospital scenario with three institutions. We evaluated models' performance in terms of fidelity, diversity, speed of training, and predictive ability of classifiers trained on the generated synthetic data. In addition, we provided explainability focused on both global and local features. Calculated distance between real images and their projections in the latent space proved the authenticity of generated samples, which is one of the main concerns in this type of applications. The code for studies is publicly available (https://github.com/aidotse/stylegan2-ada-pytorch).

Keywords: GAN · federated learning · skin lesion classification · XAI

1 Introduction

In recent years, the use of neural networks has become a very popular and attractive topic for many medical researches [7,9,17], as one of the key promises of using Artificial Intelligence (AI) in healthcare is its potential to improve diagnosis. However, to create reliable deep learning (DL) algorithms that can identify complex patterns of medical conditions, they must be trained on a large amount of data. In addition, it is desirable for the model to have a diverse range of cases, as data from a single source may be biased by the acquisition protocol or the population [6,20].

Unfortunately, preparation and annotation of medical data is a costly procedure that demands the assistance of medical specialists. Additionally, access to medical data requires a lengthy approval process due to patient privacy concerns. This makes it almost impossible for different institutions to share data and thus expertise with one another. Although there are some high quality open access dataset initiatives [9,17], there is still a great need for much more diverse and complex databases to effectively apply DL.

Synthetic data appears to be a good solution to mitigate the issues with privacy policies. It can be used in two ways - firstly as extensions of small and unbalanced datasets

S. C. Limeros and S. Majchrowska—These authors contributed equally to this work.

C. Biele et al. (Eds.): MIDI 2022, LNNS 710, pp. 93–102, 2023.
https://doi.org/10.1007/978-3-031-37649-8_10

(e.g., of rare diseases) and secondly for anonymization purposes (to replace instead of augment real samples). In both scenarios, synthetic medical data must accomplish two competing goals. The data should accurately reflect the real data and simultaneously offer strong privacy protection for the individuals whose records were used to create it.

In this work we perform a detailed study of GAN-based artificial data generation in the case of International Skin Imaging Collaboration (ISIC) 2020 [17] database using StyleGAN2-ADA [11] in conditional and unconditional settings. All trained models are evaluated in terms of both fidelity and diversity. Furthermore, we conduct an extensive latent space analysis of the generated images to better understand the structure of the real and synthetic images for the subsequent binary classification task (benign and malignant). Performed evaluations base on image editing in latent space, local and global explanations of trained classifiers. As far as we know, such detailed analysis has not been attempted before.

Moreover, to deal with a more realistic scenario where a single hospital does not have a sufficiently large dataset to generate artificial data, we simulate a scenario with three hospitals with a different amount of data each. We propose to use Federated Learning (FL) [16] with the aim to synthesise a more complex, fair and diverse dataset through the collaboration of multiple medical institutions without exchanging local data samples.

2 Materials and Methods

2.1 International Skin Imaging Collaboration Database

In our experiments, the reference dataset for real images is based on the training set of the ISIC 2020 challenge [17] extended by malignant cases from previous years' competitions [15]. The database consists of the 37 648 images – the whole ISIC 2020 dataset, adding 4522 malignant samples from ISIC 2019 – where 20% were used for validation in first phase of central trainings. Later, we splitted the training subset based on patient ID attributes. To make the FL setup more appropriate, we ensured that the data from an individual patient would not be present on more than one client. For this setup, we created 3 clients and for them, data subsets with 2k, 12k, 20k images respectively. For each client the proportion of malignant and benign was roughly the same as in the whole dataset. In all experiments, we resized the input images to 256×256 pixels.

2.2 Training Details

We investigated StyleGAN2-ADA performance using an original implementation from NVIDIA Research group[1]. We trained StyleGAN2-ADA models with each of the two classes of training set as input, as well as in a conditional setting with and without augmentations. To select the best model, we considered both the Fréchet Inception Distance (FID) [8] and Kernel Inception Distance (KID) [5] metrics, along with training speed, similarly as proposed in [4]. The classification task was performed using EfficientNet-B2 model [19], pretrained on ImageNet, with Ross Wightman's implementation[2]. During training, we used the Adam method for optimizing the network weights with an

[1] https://github.com/NVlabs/stylegan2-ada-pytorch.
[2] https://github.com/rwightman/efficientdet-pytorch.

adaptive learning rate initialized to 5×10^{-4}. We trained the models for a maximum of 20 epochs and an early stopping with a patience of 3 epochs. We applied standard data augmentation techniques, such as random rotation, horizontal and vertical flip, during the training phase for all experiments. For the experiments in a FL setup, we used Flower framework [3]. In our simulated setup, we created a network with 3 clients with different amounts of data and a server, where the weights of the trained model were exchanged every 100 iterations. We used the Federated Average (FedAvg) algorithm [14] as it is an effective and simple method that is commonly used for federated aggregation.

2.3 Evaluation Protocol

Various dimensions should be considered when evaluating GANs [2]. Firstly, fidelity as a measure of reliability, and diversity as a measure of fairness. FID and KID metrics evaluate these two characteristics, but rely on a preexisting classifier trained on ImageNet, and are insensitive to the global structure of the data distribution. Also Precision (P) and Recall (R) scores measure, respectively, the fraction of synthetic samples that look realistic (fidelity) and the fraction of real samples that the model can synthesize (diversity). Perceptual Path Length (PPL) [12] estimates whether and how much latent space is entangled or regularized, ultimately being able to capture the coherence of images. Another dimension to look at is predictive performance, referring to the fact that samples should be as useful as real data when used for the same predictive purpose. Here, we built a melanoma classifier using synthetic data for training and real data for testing. Since privacy is the most important factor in medical study, we evaluated the generalization or authenticity of the generative process [2], which measures the model capability to creation of new samples. Additionally, a survey was conducted in which experts assessed whether each of the 200 tested images is real or generated artificially by cGAN. Finally, we investigated whether it is possible to edit the image by manipulating the latent input of the trained GAN. The semantic factorization (SeFa) [18] method, as it do not need a large sample of latent vectors and auxiliary classifier, was tested to see if we could obtain directions in latent space, where the influence of one feature could be controlled while preserving the rest of the image.

3 Results

3.1 GANs Trainings

In the first phase of our experiments, we established the best model in terms of fidelity and diversity using well-known metrics such as KID, FID, P, R, and PPL (see Table 1). It is worth noting that the GAN responsible only for malignant melanoma generation (mal-GAN) had around 6 times less data than for benign cases (ben-GAN). In general, the unconditional models have lower PPL scores, showing better regularity of latent space due to the fact that they model only the distribution of one class. Additionally, the vast majority of malignant melanoma examples in ISIC 2020 and ISIC 2019 show a black dermatoscope frame, which leads to the generation of darker images.

The conditional setting was used to provide the model with a wider variety of images, since there are a lot of characteristics that are common for both classes. Achieved higher FID and KID scores confirmed that this is beneficial for the minority class (malignant). For this setting, we used ADA mechanism with and without ($_{w/o\,col}$) color augmentation, and achieved the best scores for second one. Color augmentation leads to leakage of color hue (unnatural red or violet) to the generated examples. Subjective assessment based on four responses in a qualitative survey, from two dermatologists and two deep learning experts, achieved an overall average accuracy of 54% for participants (at level 58% for dermatologists and 50% – deep learning experts). There was no feature in any image that clearly suggested to the participants that the image is either real or synthetic.

Table 1. Calculated metrics for each of the generative models tested in the centralized setting.

Scenario	KID (%)	FID	P	R	PPL
ben-GAN	0.42	7.99	**0.77**	**0.45**	60
mal-GAN	0.47	15.46	0.62	0.40	**51**
cGAN	0.32	7.33	0.75	0.42	193
cGAN$_{w/o\,col}$	**0.24**	**7.02**	0.75	0.44	101

In case of simulated hospital scenario in FL setup, we observed faster convergence (1.6 times) and improved quality of the generated images mainly for the client with the smallest data resources. As the data distributions between different clients only differed in size, we put more emphasis on the classification task with centrally trained models.

3.2 Predictive Performance with Classifier

After the evaluation with general metrics, we performed a study on predictive performance to measure how useful the synthetic data is for the subsequent task, i.e. malignant melanoma diagnosis. As a baseline for the experiments, we first train the classifier on training subset of the real images of ISIC dataset, and then tested it on the validation set. Secondly, GAN-based augmentation was performed using two types of GANs models with two scenarios: training on balanced synthetic dataset with 55k images (*syn*) and testing on real validation subset (the same as in baseline experiment) and training on real images adding 22k synthetic melanoma samples (*aug*) to balance the dataset. The introduction of highly underrepresented malignant melanoma cases improves the classification accuracy roughly of few pp. in both scenarios, as sumarised in Table 2. Overall GAN-based augmentation technique does not provide reliable improvements in case of classification using the whole ISIC 2020 and malignant samples from ISIC 2019.

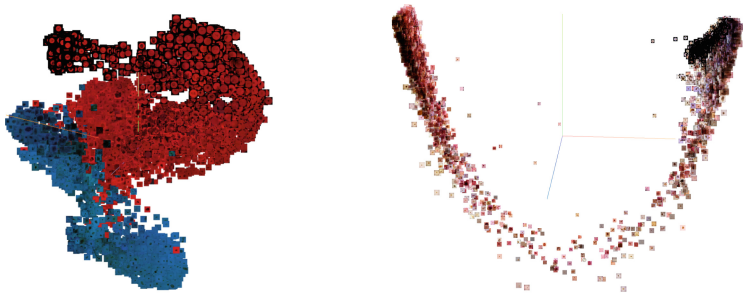
3.3 Explanations of the Predictions

To measure the authenticity we projected 12k samples from the real dataset into the latent space of the generator. This gave us the latent codes that caused our generator to synthesize the most similar output to the input image. To optimize for a latent

Table 2. Calculated metrics for each of the classification scenarios with EfficintNet-B2: trained on real (real-baseline), only synthetic samples (*syn*), and augmented balanced dataset with additional 22k fake malignant lesions images (*aug*) from conditional (cGAN) and unconditional (GAN) models. In all scenarios the models were tested on the same real images validation set.

Scenario	Acc (%)	AUC (%)	Scenario	Acc (%)	AUC (%)	Scenario	Acc (%)	AUC (%)	Scenario	Acc (%)	AUC (%)
real baseline	97.8	98.8	syn-GAN	94.1	94.2	syn-cGAN	94.7	96.7	syn-cGAN$_{w/o\,col}$	92.6	92.7
			aug-GAN	97.8	98.6	aug-cGAN	97.8	98.6	aug-cGAN$_{w/o\,col}$	**97.9**	**98.8**

code for the given input images, we followed [1]. We used a VGG16 model as a feature extractor, computed the loss on the difference of the extracted features for both the target image and the generated output, and performed backpropagation. Next, we extracted the features of both the real and their projected images using the last convolutional layer of our classifier trained on real and synthetic data (aug-cGAN$_{w/o\,col}$). These embeddings were visualized in a 3D space using t-distributed stochastic neighbor embedding (t-SNE) method [13]. This allows visually exploring the closest near neighbors of each real image using cosine distances. Figure 1 shows examples of real images projections in the latent space of the generator (with benign marked on red, malignant – blue) and projected embeddings of real and synthetic data. In both cases there is visible separation between two clusters created by two examined skin lesion classes. However, there are still plenty of the cases in the middle between two clusters and mixed with improper class, what is visible in Fig. 1(a). Additionally, we spotted some clusters inside classes, which are associated with instrumental bias, such as ruler and black dermatoscope frame.



(a) (b)

Fig. 1. Real images projections in the latent space of the generator (a). Projected embeddings of real and synthetic data coming from the classifier trained on synthetic data (b).

For a more systematic inspection, we computed the cosine distances between the different pairs of real images and their projected samples. The mean distance was equal to 0.1444 and the median 0.00283 with only two projections being *too close* in terms

of $Q1 = 0.013$ (range of $1e{-}5$) to the real images. Only in these two cases the closest neighbor was the projection of the target image, meaning that the generative model could have memorized that sample. We treated this as a measure of the authenticity of the generated samples. We also spotted that some of the images were very distant from their projections (around 2) but still resembled the target image (Fig. 2(a)).

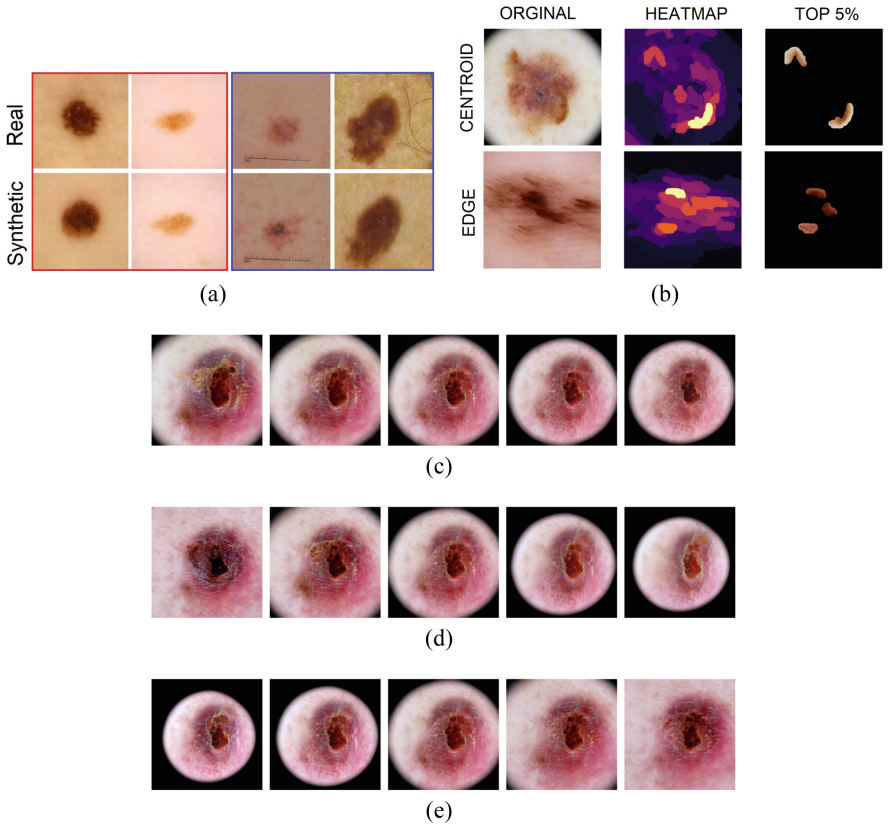


Fig. 2. A few examples of the closest (red frame) and the most distant (blue frame) pairs real-synthetic in terms of cosine distance (a). Two examples from the malignant class, which were found in the center, and in the boundary between two clusters respectively, examined using XRAI heatmaps (b). Examples of image editing using the SeFa framework shifted along the 2nd (c), 4th (d) and 6th (e) eigenvectors.

The images in the center and boundary between the two clusters (Fig. 1(b)) were studied using local explanations with the XRAI method [10]. For images of malignant lesions that belong to the centroid of the embeddings, we found that the mole itself is the most important part of the image for the final prediction. In the sample image, the network focuses on boundary pixels which represent asymmetry in the mole, one of the main clues for detecting malignant melanomas. On the other hand, in edge cases the

results are not as evident due to image distortions or poorly centred moles (Fig. 2(b)). We selected all the misclassifications and edge cases and generated N neighbors using a distance of 0.1 to augment the dataset with more complex examples with the aim of making it more robust. First experiments showed an improvement in performance in those edge cases.

Finally, we edited the latent input in an attempt to eliminate the dermoscopic frame in malignant melanoma images using the SeFa method [18]. The latent $\mathbf{w}$-vector corresponding to the image in Fig. 2(c)–2(e) is shifted along the 2nd, 4th and 6th eigenvectors. The image in the middle in all three rows (3rd column) is the original image. Left and right of the original image are positive and negative directions along these eigenvectors. The eigenvectors displayed were chosen from a larger qualitative evaluation of 100 images along the first 10 largest eigenvectors. Applying SeFa image editing suggests less entangled features from visual inspection of the images of different directions – the black frame was removed leaving the other features (such as shape, size, color) almost intact.

To assess the quality of the edited images, we first generated a large sample size of images all containing frames. After acquiring the images we removed the frames by shifting the latent vectors along the direction where the presence of the dermoscopic frame was minimized. Finally, we trained a classifier on these images for the malignant melanoma with a training set of 10k images per class and a test set consisting of real images, which result in accuracy equalled 87%.

4 Discussion

In our study, we explored the state of the art DL-based techniques to generate, classify, and explain computed results for skin lesion diagnosis. Our experiments are based on ISIC 2020 and ISIC 2019 datasets, which are one of the largest but very unbalanced open access database.

Samples generated using different types of GANs and settings exhibits slightly different appearance, as evidenced by the calculated metrics shown in the results (see Sect. 3). The PPL measure, which is capable of capturing the consistency of the images, is the lowest for generated malignant melanoma samples by unconditional GAN. However, this is not connected with the lowest KID and FID scores indicating the dissimilarity between two probability distributions (real and fake) using samples drawn independently from each distribution. Lower PPL score is related to the smallest amount of malignant data, and in result more regularized and narrow distribution of latent space. The second observation may be connected with the fact, that KID and FID rely on a pre-existing classifier (InceptionNet) trained on ImageNet that consists of different images rather than skin samples. The results also indicates that the cGAN model is prone to generating more realistic looking melanoma (using some features from benign samples) than the mal-GAN. No statistical conclusion can be drawn from the small sample size in the survey where cGAN generated images were used. However, the results do suggest that subjectively, experts are unable to tell an artificial lesion from that of a real patient. There was no specific feature that the experts picked up on in the generated data as an artifact of the model. Therefore, qualitatively the synthetic data pass for real in the eyes of experts.

In the case of classification, we have not observed a large improvement of the performance of the classification network based on synthetic data generated by StyleGAN2-ADA. Actually, the results achieved in different scenarios do not differ much. This may be affected by the large size of the real dataset, but also by the fact that some features coupled with, for example, methods of collecting data (existence of black dermoscopic frame) may be entangled with a specific class.

Performed exploration of the latent space showed that there is a clear separation between the projections of the real and generated samples. Measured distance between the projections of real and the closest synthetic image proved the authenticity of the generated samples. Our main interest in the explanation of classification results focused on the edge cases, as the dermatologists are paying special attention to those cases that lie in the boundary and are not so obvious. We noticed that the network output is often biased by acquisition protocols, as well as some patient-related features. The main issue seems to be the area covered by the mole on the image. However, this topic requires closer examination. Editing images using latent directions could be a useful tool in removing unwanted artifacts from images. Nevertheless, dermoscopic frames were present mostly in images of malignant melanoma, thus the characterization of class labels was entangled with dermoscopic frames. This entanglement resulted in changes in separate features when removing the frame artifact and did not leave the malignant melanoma data intact. For future steps, using this technique may show promising results in data normalization and generalization in different domains.

On the other hand, as GAN training requires a large investment in computing and data resources, the FL setup may be a solution for smaller institutions with a lack of access to sufficient data resources. Achieved results confirmed that generation of skin lesions in a distributed setup can lead to similar performance with respect to the quality and diversity of generated samples, with a significant faster convergence. However, to reach a final verdict on this matter, it is necessary to conduct further research into different aggregation algorithms, privacy preserving techniques, and even defense mechanisms against adversarial attacks.

5 Conclusions

GAN-based augmentation is an extensively explored technique for medical imaging applications, especially in the case of very rare diseases. First of all, it helps in the creation of larger and more balanced datasets. Secondly, it creates non-real data, which can be more easily shared amongst the medical community. However, the results achieved with the addition of synthetic data reported in literature show an improvement in accuracy of only a few percents without clearly explaining the reason. On the other hand, GAN-based anonymization suffers from an unset gold standard in measuring its performance.

To utilize GANs in generating synthetic healthcare data, a number of considerations need to be made. First, one should consider the architecture. In our case, we chose between central unconditional GANs per class, conditional GAN and FL setup. The usefulness of chosen architectures mainly depends on computational resources and time - unconditional GAN can be good option with small amount of classes due to long

duration of training of single GAN. If a massive, annotated dataset exists, training the GAN centrally is preferable but in case of a more realistic scenario of data being siloed in an institution, the benefit from FL is noticeable particularly for smaller institutions.

Second, the created synthetic data should be inspected from multiple different points of view. Common features to emphasise are fidelity and diversity, which are important to understand how well the synthetic data represents the underlying real data. Importantly, as the goal in healthcare is to avoid sharing data, it is also crucial to inspect the authenticity of the synthetic examples to make sure they are not simply copying the training data. Additionally, the synthetic data should be as useful as the real data for the subsequent task (e.g. classification) and not allow inferences based on features that are not related to the case, but, for example, to the way the data were collected (e.g., linking a black dermatoscope to malignant melanoma).

Acknowledgements. This work has been carried out during the *Eye for AI* and *Master Thesis* programs thanks to the support of Sahlgrenska University Hospital, Chalmers University of Technology, and AI Sweden.

References

1. Abdal, R., Qin, Y., Wonka, P.: Image2StyleGAN: how to embed images into the StyleGAN latent space? In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 4431–4440 (2019). https://doi.org/10.1109/ICCV.2019.00453
2. Alaa, A.M., van Breugel, B., Saveliev, E., van der Schaar, M.: How faithful is your synthetic data? Sample-level metrics for evaluating and auditing generative models. arXiv preprint arXiv:2102.08921 (2021)
3. Beutel, D.J., Topal, T., Mathur, A., Qiu, X., Parcollet, T., Lane, N.D.: Flower: a friendly federated learning research framework. arXiv preprint arXiv:2007.14390 (2020)
4. Bissoto, A., Valle, E., Avila, S.: GAN-based data augmentation and anonymization for skin-lesion analysis: a critical review. In: 2021 IEEE/CVF CVPRW, pp. 1847–1856 (2021)
5. Bińkowski, M., Sutherland, D.J., Arbel, M., Gretton, A.: Demystifying MMD GANs. arXiv preprint arXiv:1801.01401 (2021)
6. Cassidy, B., Kendrick, C., Brodzicki, A., Jaworek-Korjakowska, J., Yap, M.H.: Analysis of the ISIC image datasets: usage, benchmarks and recommendations. Med. Image Anal. **75**, 102305 (2022)
7. Gillstedt, M., Hedlund, E., Paoli, J., Polesie, S.: Discrimination between invasive and in situ melanomas using a convolutional neural network. J. Am. Acad. Dermatol. **86**(3), 647–649 (2022)
8. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In: Guyon, I., et al. (eds.) Advances in Neural Information Processing Systems, vol. 30. Curran Associates, Inc. (2017)
9. Johnson, A., Bulgarelli, L., Pollard, T., Horng, S., Celi, L.A., Roger, M.: MIMIC-IV. PhysioNet (2020)
10. Kapishnikov, A., Bolukbasi, T., Vi'egas, F., Terry, M.: XRAI: better attributions through regions. In: 2019 IEEE/CVF ICCV, pp. 4947–4956 (2019)
11. Karras, T., Aittala, M., Hellsten, J., Laine, S., Lehtinen, J., Aila, T.: Training generative adversarial networks with limited data. arXiv preprint arXiv:2006.06676 (2020)
12. Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., Aila, T.: Analyzing and improving the image quality of StyleGAN. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 8107–8116 (2020)

13. van der Maaten, L., Hinton, G.: Visualizing data using t-SNE. J. Mach. Learn. Res. **9**(86), 2579–2605 (2008)
14. McMahan, B., Moore, E., Ramage, D., Hampson, S., Arcas, B.A.Y.: Communication-efficient learning of deep networks from decentralized data. In: Singh, A., Zhu, J. (eds.) Proceedings of the 20th AIStats. Proceedings of Machine Learning Research, vol. 54, pp. 1273–1282. PMLR (2017)
15. Nozdrin, R.: Melanoma external malignant 256 (2020)
16. Rajotte, J.F., et al.: Reducing bias and increasing utility by federated generative modeling of medical images using a centralized adversary. arXiv preprint arXiv:2101.07235 (2021)
17. Rotemberg, V., et al.: A patient-centric dataset of images and metadata for identifying melanomas using clinical context. Sci. Data **8**(1), 34 (2021)
18. Shen, Y., Zhou, B.: Closed-form factorization of latent semantics in GANs. In: CVPR, pp. 1532–1540 (2021)
19. Tan, M., Le, Q.: EfficientNet: rethinking model scaling for convolutional neural networks. In: Chaudhuri, K., Salakhutdinov, R. (eds.) Proceedings of the 36th ICML. Proceedings of Machine Learning Research, vol. 97, pp. 6105–6114. PMLR (2019)
20. Wachinger, C., Rieckmann, A., Pölsterl, S.: Detect and correct bias in multi-site neuroimaging datasets. Med. Image Anal. **67**, 101879 (2021)