Published in *Proceedings of the International Conference on Electrical, Computer, Communications and Mechatronics Engineering 2022*, Maldives, Maldives, 16-18 november 2022, which should be cited to refer to this work.
DOI: 10.1109/PTC.2013.6652424

# Investigating on Data Augmentation and Generative Adversarial Networks (GANs) for Diabetic Retinopathy

Maleika Heenaye-Mamode Khan
*Faculty of Information, Communication and Digital Technologies*
University of Mauritius
Réduit 80837, Mauritius
m.mamodekhan@uom.ac.mu

Zahra Mungloo-Dilmohamud
*Faculty of Information, Communication and Digital Technologies*
University of Mauritius
Réduit 80837, Mauritius
z.mungloo@uom.ac.mu

Khadiime Jhumka
*Faculty of Information, Communication and Digital Technologies*
University of Mauritius
Réduit 80837, Mauritius
khadhiime@gmail.com

Noorshad Z. Mungloo
*Victoria Hospital*
*Ministry of Health and Wellness*
Quatre Bornes 72259, Mauritius
noorshad@hotmail.com

Carlos Peña-Reyes
*School of Management and Engineering Vaud (HES-SO)*
*University of Applied Sciences and Arts Western Switzerland Vaud*
1400 Yverdon-les-Bains, Switzerland
carlos.pena@heig-vd.ch

***Abstract*—The early detection of diabetic retinopathy (DR) disease, which is a complication of diabetes, may help to reduce blindness. One of the issues faced currently is the limited number of ophthalmologists available to take care of the diabetes patients. To overcome this problem, automated DR grading applications are required. Many researchers are now developing deep learning to build these decision support systems. Nevertheless, the frequent misclassification of DR remains an open challenge. Class imbalance in the datasets and limited labelled datasets are the root causes of misclassification. In this work, we are investigating data augmentation and Generative Adversarial Networks (GANs), known to countermeasure class imbalance and the dearth of labelled images. We have applied these two techniques on the APTOS dataset after the undersampling process. Traditional data augmentation has achieved an accuracy of 72% versus GAN which has reached an accuracy of 76%.**

***Keywords—diabetic retinopathy, convolution neural network, data augmentation, undersampling, GAN model***

## I. INTRODUCTION

Diabetic Retinopathy (DR) is an eye disease which damages the retina and may lead to complete blindness. To overcome this problem, the early detection of DR is of utmost importance[1]. DR is the consequence of untreated diabetes for a long period of time. Several automatic detection and classification applications have been developed so far. However, the correct classification of the 5 classes remains a big challenge[2].

When using supervised learning to categorise images, one of the key issues in the field of medical imaging is dealing with tiny data sets or imbalanced datasets. Several researchers have applied data augmentation to handle this challenge, including rotating, cropping, and resizing the dataset images. Using redundant data to assist network training has become a standard procedure for computer vision applications, with GANs being one of the most often utilised methods recently. In 2014, Goodfellow et al.[3] devised GANs, a form of machine learning network. The main goal of GAN is to create data that looks and behaves like genuine data, making it impossible for a human or machine observer to tell the difference. This method learns how to generate new data that has the same statistical qualities as the training batch. GAN is composed of two independent neural networks, namely, the generator and the discriminator, that constantly compete with each other. The generator tries to generate seemingly real data examples. The main objective of the discriminator is to distinguish between real and false data, that is, to separate the generated image from the real image as much as possible. The application of GANs in few other applications are bringing promising results and thus, motivating its exploration in the field of diabetic retinopathy.

In this work, the use of undersampling and GAN to handle small and imbalanced Diabetic Retinopathy fundus images datasets were explored. The aim was to investigate how these 2 techniques impacted on the accuracy of the classification model as compared to the original dataset.

The paper is organised as follows: Section II presents the related works done in diabetic retinopathy and the effect of data augmentation; Section III describes the methods and materials used in this study; in Section IV, we present the results of the study and Section V concludes the study.

## II. RELATED WORKS

For many years, machine learning techniques have provided satisfactory results in the detection and classification of diabetic retinopathy images, which eventually helped in the diagnosis of diabetes [4, 5]. In the recent few years, deep learning classifiers are gaining more attention for diabetic retinopathy classification and have been found to improve the likelihood of early diagnosis of diabetic retinopathy. The earlier trend was to adopt the existing deep learning pre-trained models, which have already been evaluated on very large datasets.

Shaban et al. [6] proposed their own Convolution Neural Network (CNN) model and achieved an accuracy of 88%-89% using the APTOS 2019 Kaggle Competition dataset.

*"© 2022 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works."*

The dataset described above shows a big difference in the different classes. Class 3 and 4 images represent only 5% and 8% respectively. To obtain a balanced dataset, before the introduction of GAN models, basic data augmentation methods were used. Methods like vertical and horizontal flip, rotation and random brightness were applied to the same class images to synthesise new images. Islam et al.[7] improved the dataset using horizontal flip, random brightness and contrast and random saturation. They proposed an InceptionV3 model and saw an increase of 5% in the test accuracy compared to the dataset without data augmentation. Meanwhile Saini et al.[8] have proposed different pre-trained models for classifying the same dataset and obtained the highest accuracy of 0.7986 with DenseNet201.

With the emergence of GAN models, some researchers have synthesized retinal fundus images using models similar to GANs to offset the imbalanced dataset. Araujo et al. [9] adopted a heuristic-based data augmentation scheme based on the synthesis of neovessel (NV)-like structures that compensates for the imbalanced Diabetic retinopathy labelled datasets. However, there was not much difference in the Quadratic weighted kappa score with and without the data augmentation. In Zhao et al. [10], Tub-sGAN was presented as a way to extend style transfer to the generator, increasing the variety of generated samples. Despite some progress, it is unable to synthesise the DR-related lesions and physiological retina features clearly. More recently, Zhou et al. [11] created high-resolution fundus images which can be manipulated with arbitrary grading and lesion information. They successfully synthesised 50 000 high quality fundus images (resolution of 1280 x 1280) paying attention to the small details. Using the pretrained models such as VGG16, ResNet50 and InceptionV3, they achieved an accuracy of 86.48%, 88.06 and 87.63 respectively with the improved dataset. With the same dataset, Ahn et al. [12] proposed their own Conditional GAN by using vessel segmentation mask and lesion segmentation mask and obtained an accuracy of 94% with the balanced dataset. Chatterjee et al. [13] proposed ForeseeGAN, a model which synthesises new images and classifies them with an accuracy of 95.9%.

There is not much of a difference in the overall accuracy when using or when not using data augmentation techniques. However, the accuracy per class has improved. Without the improved dataset, the model could differentiate from a person who has diabetic retinopathy and who has not. With the data augmentation and now with GAN models, the model can replicate the same function but also will have a much higher chance of being able to determine the severity of the DR. Thus, there is scope for further exploration of GAN on DR.

## III. MATERIALS AND METHODS

### A. Architecture of the Proposed Model

Fig. 1 shows the overall architecture of the proposed model. The model consists of 4 main steps, namely, undersampling of the dataset to handle data imbalance, data augmentation, training of the customised CNN model and performance evaluation.

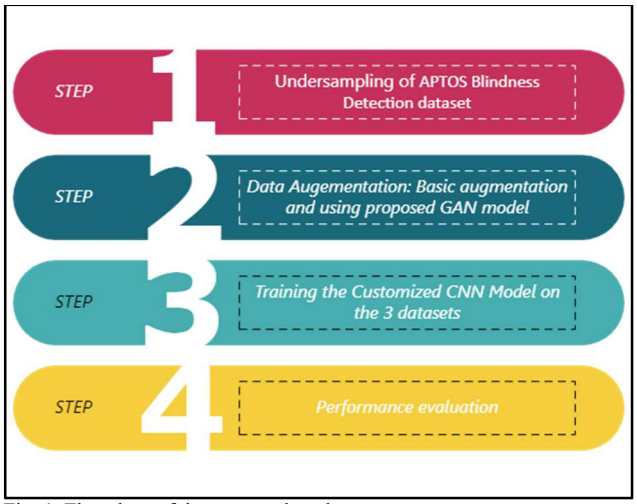


Fig. 1. Flowchart of the proposed study

### B. Datasets

The dataset used in this study is the APTOS Blindness Detection dataset available on Kaggle(https://www.kaggle.com/competitions/aptos2019-blindness-detection). The APTOS dataset is divided into 5 classes which represent the five stages of diabetic retinopathy from normal case to the proliferative stage. Fig 2 shows a detailed comparative view of the number of images from each class.

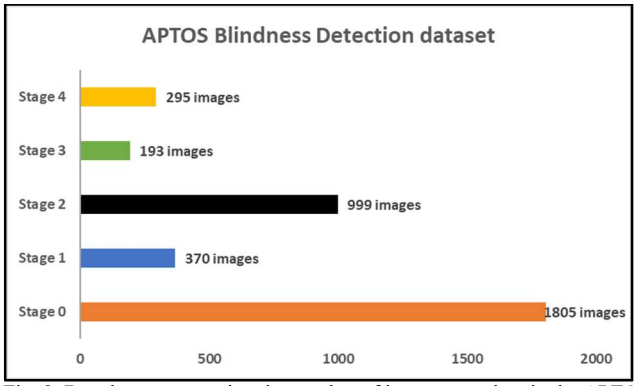


Fig. 2. Bar chart representing the number of images per class in the APTOS Blindness Detection dataset

Fig. 2. clearly reveals that the quantity of samples in each class of the 5 classes varies substantially, resulting in an unbalanced dataset. If the method is continued with this sample ratio among the classes, the classification stage may be skewed leading to overfitting or underfitting. Because the given dataset was about equally unbalanced, this problem appeared to be difficult to resolve except if we were to obtain images from more than 1000 diabetes patients suffering from different stages of diabetic retinopathy and that is very unlikely to happen. To solve the imbalance problem, undersampling is performed, that is we have taken the same number of images for each class. Since Stage 3 has the lowest, we took it as a benchmark for the other Stages and the number is a multiple of 8. This dataset is named the Original dataset. The Original dataset is divided into the ratio of 9:1, where the majority is used in the training of the model and synthesising new images, while the other one is used as a blind testing data to evaluate the performance of the model.

### C. Basic Augmentation

In this study, augmentation techniques are used to increase the number of images with the goal of improving the

performance of the model in classifying the severity of diabetic retinopathy for each image. As described in the introduction sec/tion, there will be two augmentation techniques used; basic augmentation and GAN augmentation. Before the introduction of the GAN model, basic augmentation was the main solution to increase the number of data. In the case of images, several techniques such as rotation and shearing are applied to the existing images to create new ones. In this paper, the techniques used for basic augmentation are horizontal flip, vertical flip and change in brightness. The newly constructed Basic dataset consists of 192 new images for each class adding on to the Original dataset.

### *D. GAN Model*

Another solution used to offset the class imbalance in a dataset, is to create new images using a GAN Model. In this paper, we propose a GAN model, described in the Methodology section, that is, to synthesise new images. The exact number of images is generated in order to match the same number of images in the Basic dataset. A GAN model consists of the Generator, and the Discriminator. The Generator produces new images while the Discriminator gives a probability score on how similar the new images are compared to the original images of the Original dataset, as illustrated in Fig. 3.

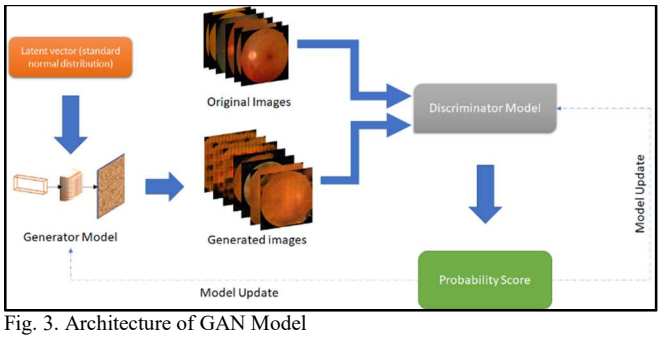


Fig. 3. Architecture of GAN Model

The proposed generator takes in a latent vector as an input, then it is passed through 6 convolutional-transpose layers, each paired with a batch normalisation layer and a ReLU activation. After passing through the first convolutional-transpose layer, which is configured with a stride of (1 x1), the latent vector becomes a block with height and width of 4 and a depth of 2048 layers. Passing through each of the following 4 convolutional-transpose layers, the depth of the building block is halved while its height and width is doubled. The last convolutional-transpose layer which is configured with a kernel size of 5 x 5 and a stride of 2 x 2, converts the output building block of the previous layer to an image similar to a fundus image with a dimension of 64 x 64 x 3. The output of the generator is fed through a Tanh function.

Meanwhile, the proposed discriminator works in a similar way as a CNN Model; it inputs the generated fundus image and outputs a scalar probability of how close the image is from the real images. This is accomplished through a series of 6 convolution layers each paired with a batch normalisation layer and a ReLU activation excluding the first and the last convolution layer. The output of the discriminator is fed through a sigmoid function to output a probability between 0 and 1. To train the discriminator, the latter is used on the original images of the APTOS dataset. The probability score tells us how similar the generated image is close to the actual images given to the discriminator. After 10 epochs, the Generator loss reaches nearly the discriminator loss and is very close to zero and the fake images are shown in a plot in Fig. 4. As a first observation, we can see that the fake images are very similar to the original images. These new images are included in the new customised GAN-synthesised dataset.

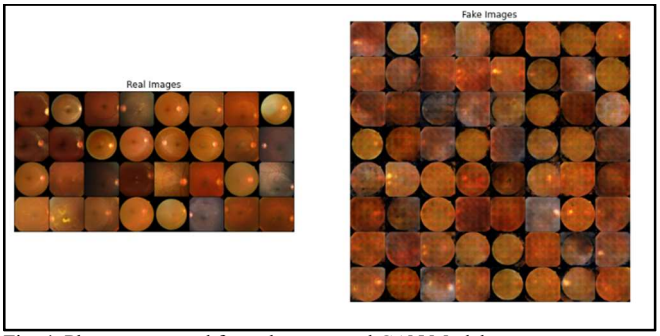


Fig. 4. Photos generated from the proposed GAN Model

### *E. CNN Model*

In this study, a customised CNN is used for the image classification of diabetic retinopathy. The CNN Model is inspired from the pre-trained VGG19 Model. The customised CNN model contains 19 layers, made up of 16 convolution layers and 2 dense layers, and its input is a 64 x 64 picture with 3 channels. The convolutional layers have a modest kernel size of 3x3 with padding and stride of 1 pixel. The model also contains 5 max-pooling layers. The activation function used in the feature extraction section is ReLU, similar to the one used in the discriminator of the GAN model. Following the extraction section, there is a classifier with two dense layers and a dropout between them, the last one has just five layers. The last dense layer is followed by a softmax layer with the same number of outputs as the previous dense layer, which yields the probability of the number of stages in diabetic retinopathy.

The network was trained using the Adam optimizer. The batch size was 8 and the number of epochs was set to 30. The dropout between the dense layers had a probability of 0.5. The learning-rate started at $10^{-5}$ and it was reduced by half each time the validation loss stopped to decrease. An early stopping function was added in case the model starts to overfit. Before the model is allowed to train on the training set, the latter is further split into training data and validation data with the ratio of 8:2.

## IV. RESULTS AND DISCUSSION

First of all, we have applied the undersampling methods, whereby some observations of the majority class were decreased. Fig. 5 shows the results obtained from the undersampling method.

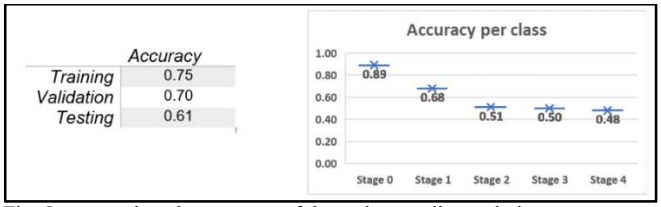


Fig. 5. summarises the accuracy of the undersampling technique

It can be observed that an overall testing accuracy of 61% was achieved for all the five classes. From the second graph, the trained model could distinguish Stage 0 and Stage 1 easier than the other Stages.

### A. *Basic Data augmentation*

Basic data augmentation was conducted on the images. Fig. 6 shows some images after the application of basic data augmentation.

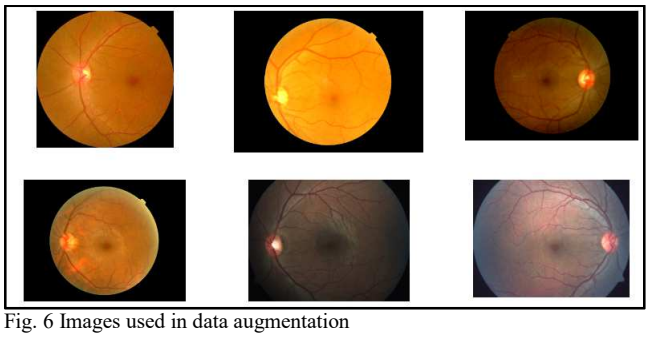

Fig. 6 Images used in data augmentation

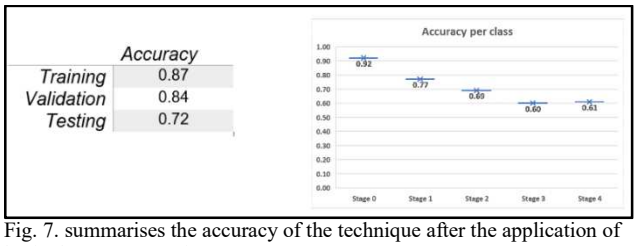


Fig. 7. summarises the accuracy of the technique after the application of basic data augmentation.

An overall accuracy of 72% was achieved. Fig. 7 displays the accuracies of each individual class as well as the training, the validation and the testing accuracies across all classes. It can be observed that basic data augmentation has a positive effect on the overall accuracy of the application. This is also confirmed by Pham et al. [14] where a better result was obtained with data augmentation compared to traditional methods for the classification of skin lesions. Likewise, in a similar domain, Lee and Chin [15] have also achieved better accuracy using data augmentation. The accuracy for each class was higher compared to that of the Original dataset.

### B. *GAN augmentation results*

GAN is yet another data augmentation method, applied on the DR datasets. Fig. 8 displays the results obtained after the application of the GAN method. On the left of the image, the training, the validation and the testing accuracy across all classes are shown and on the right, the accuracy achieved in each class is displayed.

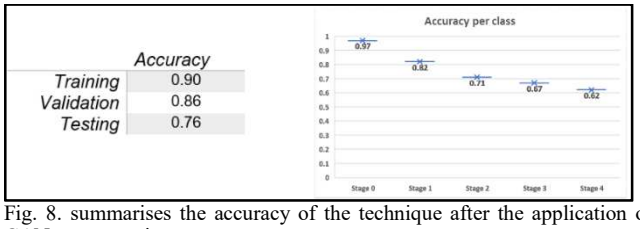


Fig. 8. summarises the accuracy of the technique after the application of GAN augmentation.

It can be observed that an overall accuracy of 76% has been achieved when using GAN compared to 72% for basic data augmentation and 61% for undersampling. This shows that the number of samples in the datasets and class imbalance impact on the accuracy of the model. In a work conducted by Qin et al. [16], GAN has improved the accuracy of the skin cancer classifier. Similar conclusions were drawn in this work, whereby GAN outperformed the other techniques. It is thus observed that the number of images of a dataset affects the performance of the classifier.

Another interesting observation was the decline in the accuracy achieved from Stage 0 to Stage 4 across the different datasets. For the 3 datasets, the highest accuracy was achieved for Stage 0, followed by, in decreasing order, Stage 1, Stage 2, Stage 3 and finally Stage 4. It shows that Stage 0 was the easiest to differentiate and stage 4 was the most challenging to classify, in all cases and even when the data was augmented and balanced. This requires further investigation.

## V. CONCLUSION

The early detection of diabetic retinopathy is of utmost importance to avoid catastrophic outcomes like blindness. Retinography or Ophthalmoscopy is the most common tool for the diagnostics of the disease. However, the lack of ophthalmologists in many hospitals has led to the need for developing computer assisted image tools to aid and improve the diagnostic decision. Deep learning has been adopted lately to develop such applications. However, deep learning needs to be trained on large labelled datasets, which are not always available in the medical field. Data augmentation and GAN have emerged as appealing techniques that can be adopted to overcome the problems of limited datasets and class imbalance. In this work, we have investigated the application of basic data undersampling, augmentation and GAN. Based on the results obtained, it can be concluded that GAN achieved better results compared to basic data augmentation and undersampling methods.


## ACKNOWLEDGMENT

This project has been funded by the Higher Education Commission (HEC) under grant number T0714. The funders had no role in the design, implementation, decision to publish or preparation of the manuscript.


## REFERENCES


[1] W. Zhang, J. Zhong, S. Yang, Z. Gao, J. Hu, Y. Chen, and Z. Yi, Automated identification and grading system of diabetic retinopathy using deep neural networks. Knowledge-Based Systems. 175, 12–25 (2019).

[2] T. R. Gadekallu, N. Khare, S. Bhattacharya, S. Singh, P. K. R. Maddikunta, I.-H. Ra, and M. Alazab, Early Detection of Diabetic Retinopathy Using PCA-Firefly Based Deep Learning Model. Electronics. 9, 274 (2020).

[3] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, Generative adversarial networks. Commun. ACM. 63, 139–144 (2014).

[4] V. Lakshminarayanan, H. Kheradfallah, A. Sarkar, and J. J. Balaji, Automated detection and diagnosis of diabetic retinopathy: A comprehensive survey. J. Imaging. 7 (2021), doi:10.3390/jimaging7090165.

[5] N. Tsiknakis, D. Theodoropoulos, G. Manikis, E. Ktistakis, O. Boutsora, A. Berto, F. Scarpa, A. Scarpa, D. I. Fotiadis, and K. Marias, Deep learning for diabetic retinopathy detection and classification based on fundus images: A review. Comput. Biol. Med. 135, 104599 (2021).

[6] M. Shaban, Z. Ogur, A. Mahmoud, A. Switala, A. Shalaby, H. Abu Khalifeh, M. Ghazal, L. Fraiwan, G. Giridharan, and H. Sandhu, A. S. El-Baz, A convolutional neural network for the screening and staging of diabetic retinopathy. PLoS ONE. 15, e0233514 (2020).

[7] N. Islam, U. Saeed, R. Naz, J. Tanveer, K. Kumar, and A. A. Shaikh, in 2019 2nd International Conference on new Trends in Computing Sciences (ICTCS) (IEEE, 2019), pp. 1–6.

[8] M. Saini, and S. Susan, Diabetic retinopathy screening using deep learning for multi-class imbalanced datasets. Comput. Biol. Med. 149, 105989 (2022).

[9] T. Araujo, G. Aresta, L. Mendonca, S. Penas, C. Maia, A. Carneiro, A. M. Mendonca, and A. Campilho, Data augmentation for improving proliferative diabetic retinopathy detection in eye fundus images. IEEE Access. 8, 182462–182474 (2020).

[10] H. Zhao, H. Li, S. Maurer-Stroh, and L. Cheng, Synthesizing retinal and neuronal images with generative adversarial nets. Med. Image Anal. 49, 14–26 (2018).

[11] Y. Zhou, B. Wang, X. He, S. Cui, and L. Shao, DR-GAN: Conditional Generative Adversarial Network for Fine-Grained Lesion Synthesis on Diabetic Retinopathy Images. IEEE J. Biomed. Health Inform. PP (2020), doi:10.1109/JBHI.2020.3045475.

[12] S. Ahn, Q. T. M. Pham, J. Shin, and S. J. Song, Future image synthesis for diabetic retinopathy based on the lesion occurrence probability. Electronics. 10, 726 (2021).

[13] S. Chatterjee, R. Mitra, and A. Dey, Diabetic retinopathy detection using 'ForeseeGAN': A deep learning approach. Open Science Framework. 65 (2022), doi:10.17605/osf.io/e37gt.

[14] T.-C. Pham, C.-M. Luong, M. Visani, and V.-D. Hoang, in Intelligent information and database systems, N. T. Nguyen, D. H. Hoang, T.-P. Hong, H. Pham, B. Trawiński, Eds. (Springer International Publishing, Cham, 2018), vol. 10752 of Lecture notes in computer science, pp. 573–582.

[15] K. W. Lee, and R. K. Y. Chin, in 2020 IEEE 2nd International Conference on Artificial Intelligence in Engineering and Technology (IICAIET) (IEEE, 2020), pp. 1–6.

[16] Z. Qin, Z. Liu, P. Zhu, and Y. Xue, A GAN-based image synthesis method for skin lesion classification. Comput. Methods Programs Biomed. 195, 105568 (2020).