Graefe's Archive for Clinical and Experimental Ophthalmology (2023) 261:1399–1412
https://doi.org/10.1007/s00417-022-05919-9

MEDICAL OPHTHALMOLOGY

Generative adversarial network-based deep learning approach in classification of retinal conditions with optical coherence tomography images

Ling-Chun Sun[1] · Shu-I. Pao[2] · Ke-Hao Huang[3] · Chih-Yuan Wei[4] · Ke-Feng Lin[5,6] · Ping-Nan Chen[7]

Received: 19 May 2022 / Revised: 21 September 2022 / Accepted: 22 November 2022 / Published online: 28 November 2022

Abstract

Purpose To determine whether a deep learning approach using generative adversarial networks (GANs) is beneficial for the classification of retinal conditions with Optical coherence tomography (OCT) images.

Methods Our study utilized 84,452 retinal OCT images obtained from a publicly available dataset (Kermany Dataset). Employing GAN, synthetic OCT images are produced to balance classes of retinal disorders. A deep learning classification model is constructed using pretrained deep neural networks (DNNs), and outcomes are evaluated using 2082 images collected from patients who visited the Department of Ophthalmology and the Department of Endocrinology and Metabolism at the Tri-service General Hospital in Taipei from January 2017 to December 2021.

Results The highest classification accuracies accomplished by deep learning machines trained on the unbalanced dataset for its training set, validation set, fivefold cross validation (CV), Kermany test set, and TSGH test set were 97.73%, 96.51%, 97.14%, 99.59%, and 81.03%, respectively. The highest classification accuracies accomplished by deep learning machines trained on the synthesis-balanced dataset for its training set, validation set, fivefold CV, Kermany test set, and TSGH test set were 98.60%, 98.41%, 98.52%, 99.38%, and 84.92%, respectively. In comparing the highest accuracies, deep learning machines trained on the synthesis-balanced dataset outperformed deep learning machines trained on the unbalanced dataset for the training set, validation set, fivefold CV, and TSGH test set.

Conclusions Overall, deep learning machines on a synthesis-balanced dataset demonstrated to be advantageous over deep learning machines trained on an unbalanced dataset for the classification of retinal conditions.

Keywords Generative adversarial networks · Imbalance · OCT · Deep learning · Synthesis-balanced

✉ Ping-Nan Chen
g931310@gmail.com; g931310@mail.ndmctsgh.edu.tw

[1] School of Medicine, National Defense Medical Center, Taipei, Taiwan

[2] Department of Ophthalmology, Tri-Service General Hospital, National Defense Medical Center, Taipei, Taiwan

[3] Department of Ophthalmology, Song-Shan Branch of Tri-Service General Hospital, National Defense Medical Center, Taipei, Taiwan

[4] Graduate Institute of Life Sciences, National Defense Medical Center, Taipei, Taiwan

[5] Medical Informatics Office, Tri-Service General Hospital, National Defense Medical Center, Taipei, Taiwan

[6] School of Public Health, National Defense Medical Center, Taipei, Taiwan

[7] Department of Biomedical Engineering, National Defense Medical Center, No.161, Sec.6, Minchiuan E. Rd., Neihu Dist, Taipei 11490, Taiwan

Key messages

- As demonstrated in several studies, underdiagnosis of retinal conditions in professional healthcare practices is common.
- Using generative adversarial networks to build synthesis-balanced datasets could foster more robust deep learning machines to aid physicians in making accurate and timely diagnoses.
- Deep learning machines trained with a synthesis-balanced dataset present an edge over machines trained with an unbalanced dataset for the classification of retinal conditions with OCT images.

Introduction

The retina processes rays of light into the vision center of our neural circuitry, thereby images of what we perceive are formed. Pathological disorders of this region beget differential changes in visual acuity, visual field defects, and even blindness [1]. Early detection of such disorders is pivotal to prevent rapid exacerbation, facilitate good prognosis, and minimize the risk of complete vision loss [2–5]. However, observational studies have demonstrated though patients undergo eye examinations in professional healthcare practices, inclusive of ophthalmologists, conditions are often underdiagnosed [6–8]; consequently, diseases could be undetected or mistreated until the patient undergoes further assessment and examination. This perplexity can be attributed to insufficient ophthalmic education received by non-ophthalmologist physicians [9] or perhaps ophthalmologists' fatigue influencing their diagnostic accuracy [10, 11]. Whatever the case, poor detection of retinal diseases is a problem; and as the incidence and prevalence of retinal diseases are projected to proliferate in countries across the world [12–16], developments are necessary to combat the mounting matter.

Optical Coherence Tomography (OCT) is the mainstay imaging technique for ophthalmic care; and is central in the diagnosis, management, and treatment of retinal disease [17–19]. With an eye towards physicians primed to take on the snowballing wave of patients with retinal problems, researchers have developed OCT image-based deep-learning techniques which could aid physicians in making more conclusive diagnoses [20–27]. Studies have demonstrated that the accuracy of deep learning machines does not pale in comparison with veteran ophthalmologists in the classification of serous macular detachment, cystoid macular edema, epiretinal membrane, macular hole, and central serous chorioretinopathy [26, 27]; all of which are retinal conditions and diseases.

Though deep learning machines yield promising results for OCT image-based detection of retinal diseases [20–27], they still bear a fundamental class imbalance issue [28]: machines trained on datasets with an uneven or skewed class distribution do not present a reliable benchmark performance for the dataset as a collective—regarding each class [29, 30]. To rectify the class imbalance dilemma, researchers have augmented their data with label-preserving transformations [31–34] (e.g., geometric, color, texture…); however, the method is generic, and the removal of certain transformations may even result in decreased accuracy of the machine [35]. In recent years, generative adversarial networks (GANs)—a technique that augments datasets specific to dataset features—were used in place of label-preserving transformations to augment and balance datasets [36–38]. Studies have substantiated the feasibility of GANs in producing realistic medical images and employing the generated images to resolve the class imbalance issue [36–38]; even so, GAN research with retinal OCT images seems to have only explored the validity of using GANs to generate high fidelity OCT images [39].

Therefore, in this study, we aim to explore the implementation of GANs in a novel deep learning-based approach for the classification of retinal OCT images. For results to be representative of current GAN developments, the record-breaking StyleGAN2-ADA [40] was chosen for our GAN model. To be compared and replicated for future experimentation, Kermany's publicly available dataset [41] was selected to train our deep learning approach. Kermany's dataset was collected from hospitals in the USA and China, and patient demographics indicate the large majority of retinas are of Caucasian origin [41]. We further test the performance of our machines, trained with heterogeneous retinas, on an external dataset obtained from the Tri-service General Hospital (TSGH) comprised of retinas predominantly of Chinese populations. The outcomes of this study may provide insight into future GAN implementations and set a standard for future retinal OCT image-based deep learning machines trained on synthesis-balanced datasets.

Methods

The experimental protocol was approved by the Tri-Service General Hospital (TSGH) human ethics committee under registration number IRB: 1–108-05–082. Images were

retrospectively obtained from the Department of Ophthalmology and the Department of Endocrinology and Metabolism at TSGH and were anonymized; thus, informed consent was not required.

Datasets

Kermany's publicly available dataset was selected to train our deep learning machines. The dataset contains a training set and test set of images (hereinafter referred to as "Kermany training set" and "Kermany test set"). The Kermany training set is categorized by condition, encompassing: Choroidal neovascularization (CNV), Diabetic macular edema (DME), DRUSEN, and NORMAL; each class consisting of 37,205, 11,348, 8616, and 26,315 images, respectively. The Kermany test set is categorized identically, however, all classes consist of 242 images. To assess the clinical feasibility of our approach, retinal OCT images were collected from 2082 patients who visited the Department of Ophthalmology and the Department of Endocrinology and Metabolism at TSGH from January 2017 to December 2021 (hereinafter referred to as "TSGH test set"); two ophthalmologists with over 5 years of clinical experience manually classified the images obtained to construct the TSGH test set. The distribution of images in datasets is summarized in Table 1.

Table 1 Distribution of images used in this study

Condition	Kermany dataset		TSGH dataset
	Kermany training set	Kermany test set	TSGH test set
CNV	37,205	242	356
DME	11,348	242	777
DRUSEN	8616	242	145
NORMAL	26,315	242	804
Total images	83,484	968	2082

The Kermany training set—constitutes the "unbalanced training set"—and is balanced in accordance with the largest class (CNV)—using StyleGAN2-ADA—and the smallest class (DRUSEN)—via down-sampling; the training set is balanced in accordance with the largest class is hereinafter referred to as the "synthesis-balanced training set," and the training set balanced in accordance with the smallest class is referred to as the "reduction-balanced training set." We trained machines utilizing the balanced and unbalanced training sets in a fivefold cross-validation framework, and outcomes were compared. Each training set was split into five fixed equal subsets. Of them, four subsets were used for training and one

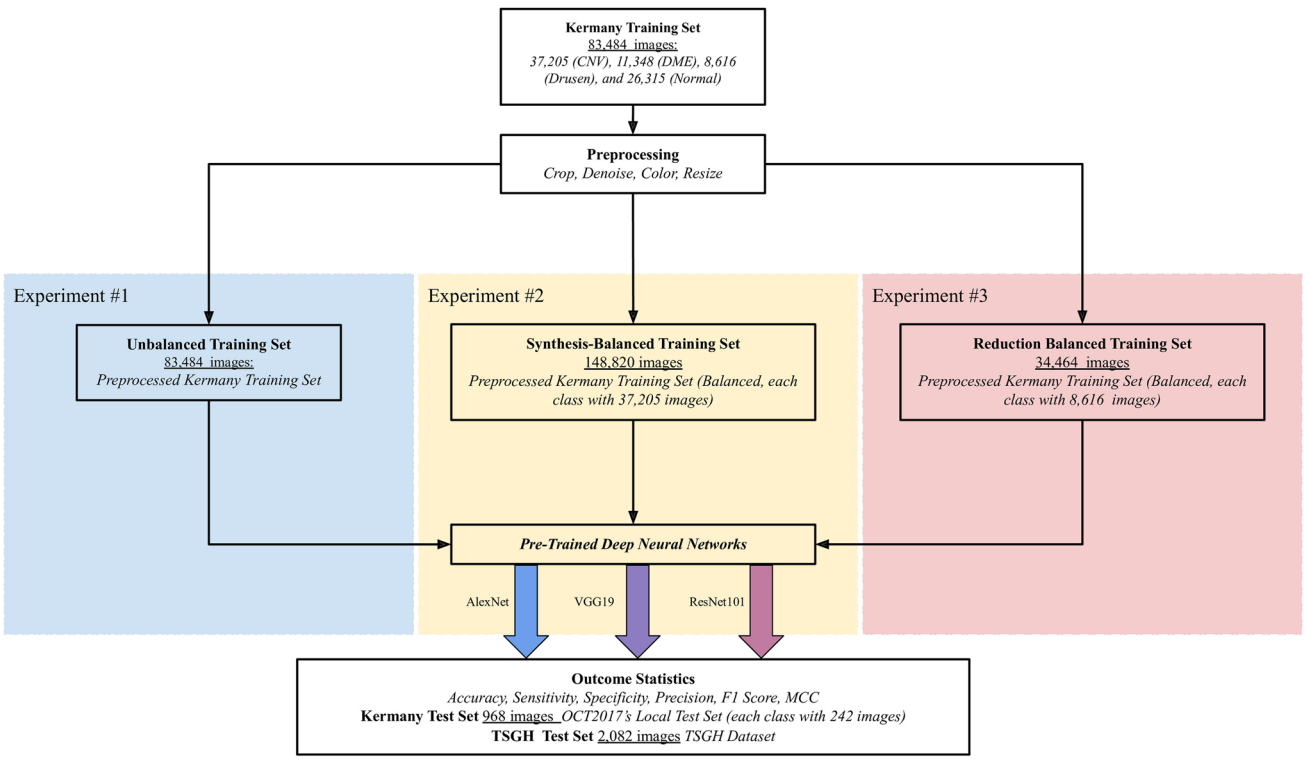


Fig. 1 Flowchart of training deep learning machines on unbalanced, synthesis-balanced and reduction-balanced OCT datasets for the classification of retinal conditions

was used for validation; this process was repeated for a total of 5 iterations, in which each iteration employed a different subset for validation. In further exploration of deep learning frameworks, the effect of differential pretrained deep neural networks (DNNs) on outcomes was also accessed. The TSGH test set was used to evaluate the performance of our deep learning machines. The flowchart (Fig. 1) below summarizes the study.

Preprocessing

There is substantial variability in the resolution of images in the Kermany training set, thus for optimized training progression, it is pivotal to normalize image resolutions. As part of the image, the white space was also resized alongside the focus of the image: the retina and so direct conversion of large resolution images (1536 × 496) into desired resolution size was not training effective and resulted in a loss of training-relevant regions. Thus, to maintain the principal and relevant areas of the image, the white space in images was zealously cropped before resizing. To denoise the images, shadows and highlights were adjusted by factors of 100 and – 100, respectively. The process resulted in images with minimized irrelevant regions (the white space) and maximized training-relevant regions. In addition, images were normalized to a color depth of 32 bits and to a resolution size of 512 × 512 to fulfill input requirements for the StyleGAN2-ADA network.

GAN architecture design

The fundamental architecture of a GAN consists of a generator (which generates fake images), real images, and a discriminator (which differentiates between real images and fake images).

The GAN implemented for this study was StyleGAN2-ADA, an improved version of the original StyleGAN developed by Karras et al. [40]. The ADA stands for adaptive discriminator augmentation and is explained further in this section. A pretrained DNNs, trained on the Flickr-Faces-HQ (FFHQ) dataset with resolution 512 × 512 by Karras et al. (FFHQ512) was employed for transfer learning; and the first 13 layers of the discriminator was frozen. StyleGAN2-ADA's augmentation pipeline consists of pixel blitting, geometric transformation, color transformation, image-space filtering, additive noise, and cutout, and all were applied in our study. The figure below (Fig. 2) outlines the architecture of the GAN network used in this study and is described hereinafter.

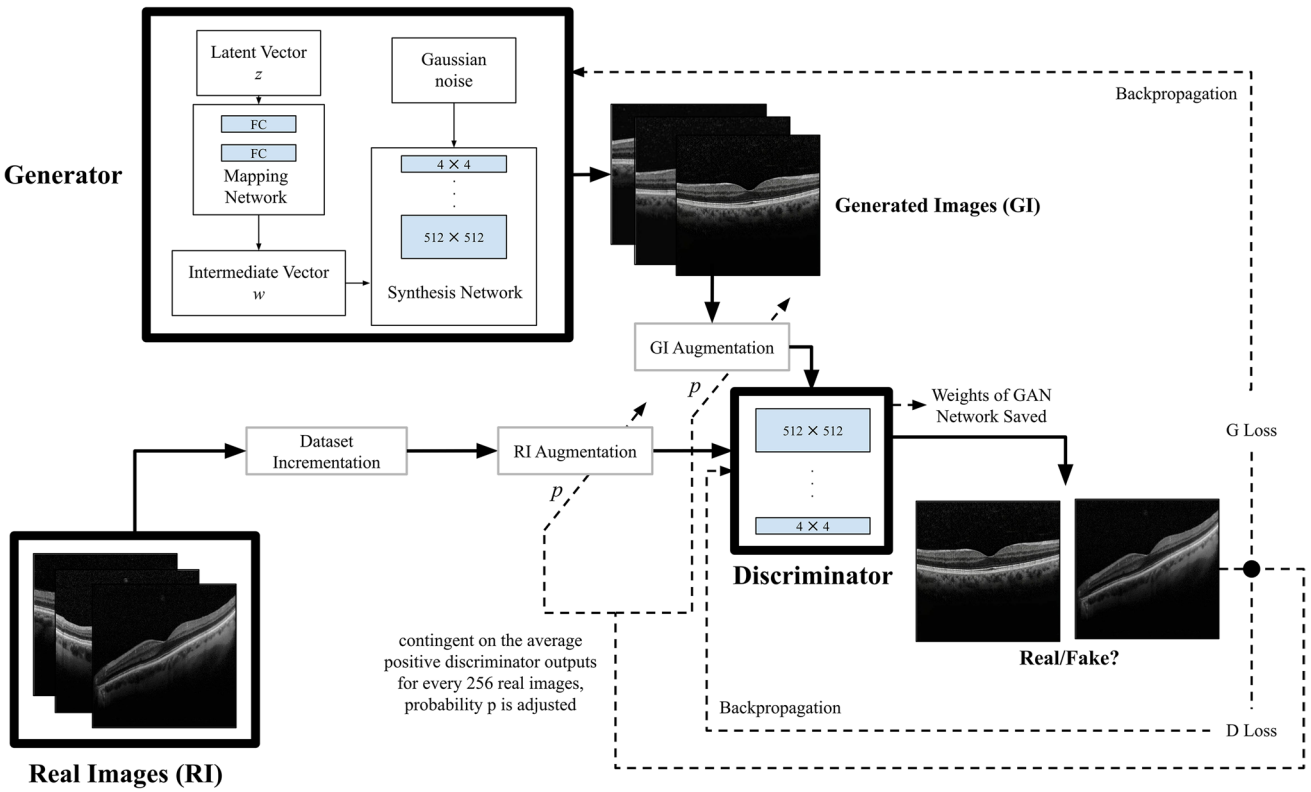


Fig. 2 GAN architecture used for this study

1) Generator: In the generator, latent vector z is normalized and mapped–via. a mapping network comprised of 2 fully connected layers–into intermediate vector w and is fed into the synthesis network—also a constitution of the generator—alongside uncorrelated gaussian noise. The synthesis network, comprised of 39 layers, generates fake images, and the images are augmented (GI augmentation) proportionally by probability p—adjusted according to the positive discriminator outputs (when the discriminator classifies images correctly) for every 256 images—before introduction to the discriminator; if probability p is 0.2, then 20% of the images are augmented.
2) Real images: The real images were normalized to 512 × 512, and the dataset is increased twofold with random flips (dataset incrementation). The dataset supplies the discriminator with real images, and the images are augmented (RI augmentation) proportionally by probability p (same *p*-value as GI augmentation) prior to introduction to the discriminator.
3) Discriminator: The discriminator—composed of 33 layers—evaluates the authenticity of real and fake images; when real images are classified as "fake" or fake images are classified as "real", the discriminator is strengthened through backpropagation from the discriminator loss (D loss) generated from the misclassifications. When fake images are classified as "fake" or real images are classified as "real" by the discriminator, generator loss is effectively generated (G loss), and through backpropagation, the generator is strengthened. The average positive discriminator output for every 256 images of the real image dataset is used to determine whether there is overfitting, and the *p*-value is adjusted accordingly. The threshold is 0.6; if the average output is above 0.6, the *p*-value is increased and vice versa. The *p*-value is systematically adjusted throughout training, hence the name adaptive, and this mechanism is specific to StyleGAN2-ADA.

With time, the generator learns to better synthesize plausible images and the discriminator learns to better distinguish real and fake images. For every 200 thousand real images shown to the discriminator, the instantaneous weights of the GAN network are saved. FID—an indication of the degree of similarity between real and fake images—is subsequently calculated. The network was trained to 1 million images shown to the discriminator. Classes were trained individually; thus, three individual networks (one for each condition) were generated.

Optimal GAN augmentation selection

Before images were balanced with StyleGAN2-ADA, a preliminary assessment of fixed order discriminator augmentations (shown in the table below), trained to 8000 images, was done to select the optimal augmentation—in terms of FID—to train our images in the GAN network. Results of the preliminary assessment are shown below (Table 2); as shown, augmentation "bgcfnc" yielded the most promising FID scores of 30.07, 33.70, and 31.12 for classes DME, DRUSEN, and NORMAL, respectively. Thus, "bgcfnc" was selected as the discriminator augmentation configuration in the training of all applicable classes.

Image synthesis

Images were synthesized with the network described previously. Images for each condition were synthesized to the largest class of images (CNV: 37,205). DME contained 11,348 images, thus 25,857 images were synthesized: DRUSEN contained 8616 images, thus 28,859 images were synthesized: NORMAL contained 26,315 images, thus 10,890 images were synthesized. After the image synthesis stage, each class contained 37,205 images, and final FID scores of 24.69, 27.35, and 22.89 were obtained for conditions, DME, DRUSEN, and NORMAL, respectively. Examples of synthesized images are shown below (Fig. 3). For retinal conditions, DME, DRUSEN, and NORMAL, the row above (A, B, C) are examples of their respective real OCT images; the row below (D, E, F) are examples of their respective synthesized OCT images.

Table 2 Performance of GAN network with different augmentations on OCT images of retinal conditions

Condition	Augmentations					
	b	bg	bgc	bgcf	bgcfn	bgcfnc
DME	37.20	139.85	56.01	34.46	33.21	30.07
DRUSEN	43.12	63.14	49.74	36.17	40.40	33.70
NORMAL	39.65	93.31	47.34	33.23	33.28	31.12

b, pixel blitting; *bg*, pixel blitting, geometric transformation; *bgc*, pixel blitting, geometric transformation, color transformation; *bgcf*, pixel blitting, geometric transformation, color transformation, image-space filtering; *bgcfn*, pixel blitting, geometric transformation, color transformation, image-space filtering, additive noise; *bgcfnc*, pixel blitting, geometric transformation, color transformation, image-space filtering, additive noise, cutout

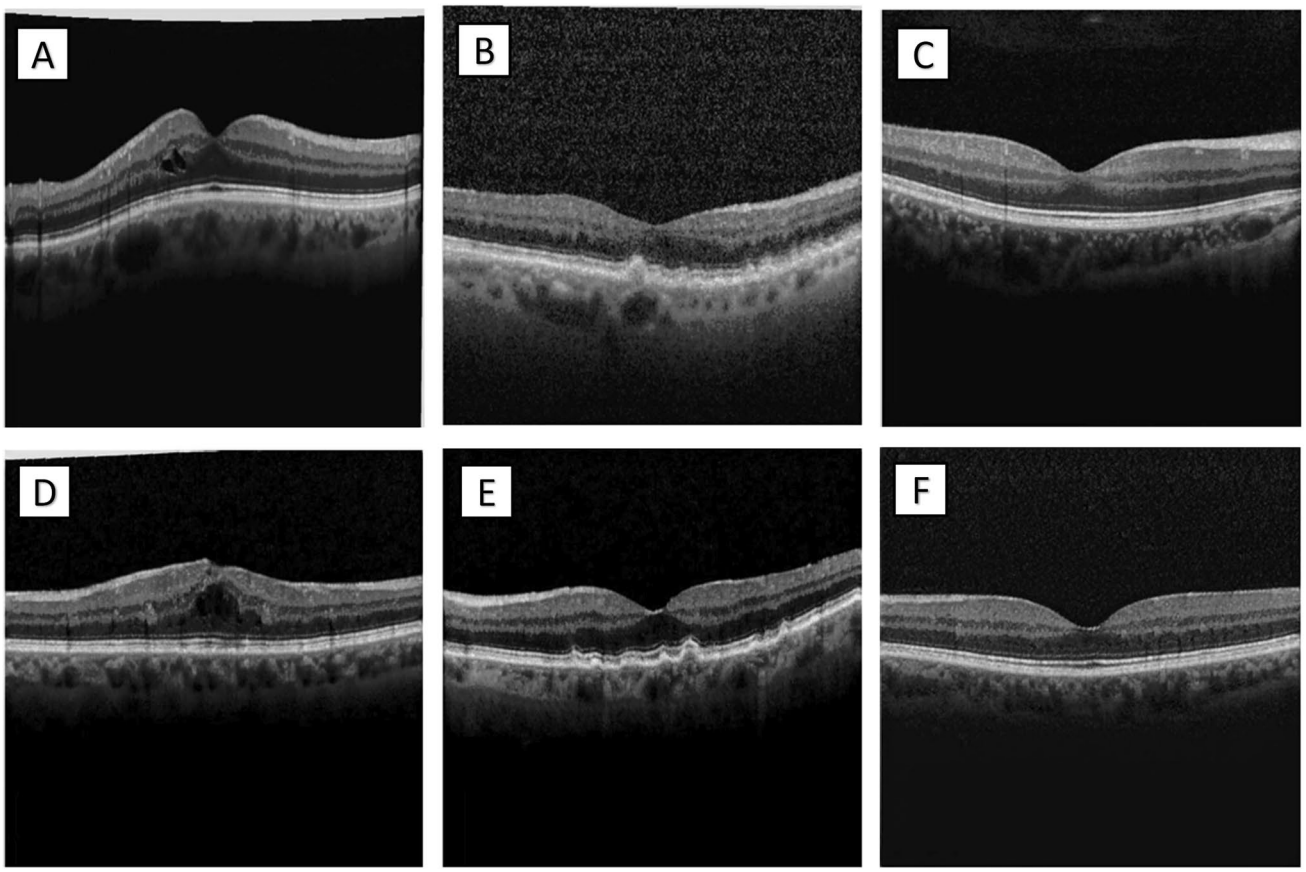


Fig. 3 Synthetic and real OCT images of retinal condition

Transfer learning with pre-trained deep neural networks (DNNs)

Transfer learning with pretrained DNNs was utilized for deep learning in our study. Machines were trained to differentiate between CNV, DME, DRUSEN, and NORMAL, and effectively classify retinal OCT images. Pretrained DNNs, AlexNet, VGG19, and ResNet101 were selected from Matlab's 2021a documentation [42] of pretrained DNNs and were employed for our study: all networks were used to train the unbalanced, synthesis-balanced, and reduction-balanced datasets. Our deep learning machines were trained with the optimization algorithm ADAM with a learning rate of 0.0001, a mini-batch size of 64, and the number of epochs was fixed at 10 epochs. In order to prevent the model from overfitting, we augmented images with "RandRotation" = (− 5, 5), "RandXReflection" = 1, "RandXShear" = (− 0.05 0.05), "RandYShear" = (− 0.05 0.05). Prior to training, the resolution size of images was normalized to fulfill respective pretrained DNN input requirements.

Evaluation methodology

To evaluate the performance of deep learning machines trained on an unbalanced, synthesis-balanced, and reduction-balanced datasets, experiments of machines trained on an unbalanced dataset (experiment #1), synthesis-balanced dataset (experiment #2), reduction-balanced dataset (experiment #3)–three experiments–were carried out and outcomes were compared.

In the three experiments, datasets were trained on selected pretrained DNNs, and their performance was judged on their capacity to correctly classify retinal OCT images on their respective training set, validation set, fivefold CV, Kermany test set, and TSGH test set. The best-performing iteration was chosen to represent each dataset.

Outcome statistics

The accuracy, sensitivity, specificity, and precision of machines on each condition (CNV, DME, DRUSEN, and NORMAL) as well as their averages were calculated. The F1 score and Matthews correlation coefficient (MCC)

Table 3 Experimental results of deep learning machines trained on an unbalanced dataset on selected pretrained DNNs in classification of retinal OCT images (experiment 1)

	Accuracy (%)	Sensitivity (%)	Specificity (%)	Precision (%)	F1 score (%)	MCC (%)
AlexNet						
Training set	97.40	96.08	99.11	96.40	96.23	95.34
Validation set	95.80	93.62	98.54	94.25	93.91	92.47
5-Fold CV	96.40	94.21	98.71	95.32	94.74	93.51
Kermany test set	99.28	99.28	99.76	99.30	99.28	99.04
TSGH test set	67.72	63.45	88.61	68.76	63.35	54.42
VGG19						
Training set	96.72	95.78	98.88	95.19	95.45	94.32
Validation set	96.01	94.72	98.62	94.32	94.47	93.10
5-Fold CV	96.41	94.24	98.68	95.51	94.85	93.61
Kermany test set	99.59	99.59	99.86	99.59	99.59	99.45
TSGH test set	79.20	73.72	92.19	77.59	75.15	68.02
ResNet101						
Training set	97.73	97.62	99.28	96.04	96.78	96.02
Validation set	96.51	95.80	98.85	94.47	95.10	93.90
5-Fold CV	97.14	96.48	99.06	95.49	95.96	94.98
Kermany test set	99.59	99.59	99.86	99.59	99.59	99.45
TSGH test set	81.03	81.18	92.94	82.26	81.28	74.85

CV cross-validation

were used to evaluate the quality of machines for the classification of retinal conditions with OCT images. Moreover, receiver operating characteristic (ROC) curve scores were also calculated to illustrate the performance of the machines.

Experimental environment

The experiments were run on Matlab 2021a Deep Learning Toolbox in the Windows 10 operating system with 2 CPUs (Intel® Xeon® Platinum 8165 Processors), 128 GB memory, and 3 graphics cards (NVIDIA® Tesla® P100).

Results

Experiment #1: Performance of deep learning machines trained on an unbalanced dataset in classification of retinal OCT images

Table 3 illustrates the statistical outcomes of training the unbalanced dataset with different pretrained deep neural networks (DNNs).

For the training set, pretrained DNN ResNet101 was the top classifier, with an accuracy of 97.73%, sensitivity of 97.62%, specificity of 99.28%, precision of 96.04%, F1 score of 96.78%, and MCC score of 96.02%.

In terms of the validation set, pretrained DNN ResNet101 was the top retinal OCT image classifier, with an accuracy of 96.51%, sensitivity of 95.80%, specificity of 98.85%, precision of 94.47%, F1 score of 95.10%, and MCC score of 93.90%.

With fivefold CV (cross validation), pretrained DNN ResNet101 was the top classifier, with an accuracy of 97.14%, sensitivity of 96.48%, specificity of 99.06%, precision of 95.49%, F1 score of 95.96%, and MCC of 94.98%.

With the Kermany test set, pretrained DNN ResNet101 was the top classifier, with an accuracy of 99.59%, sensitivity of 99.59%, specificity of 99.86%, precision of 99.59%, F1 score of 99.59%, and MCC of 99.45%.

When tested with the TSGH test set, pretrained DNN ResNet101 was the top classifier, with an accuracy of 81.03%, sensitivity of 81.18%, specificity of 92.94%, precision of 82.26%, F1 score of 81.28%, and MCC score of 74.85%.

In all sets, accuracies of DNNs increasingly progressed from AlexNet to VGG19 to ResNet101. For sets derived from the Kermany dataset, pretrained DNNs produced high classification accuracies; however, when tested with the TSGH test set, classification accuracies slumped tremendously. Moreover, ResNet101 demonstrated to be the dominant DNN as it achieved the highest accuracies for all sets.

Experiment #2: Performance of peep learning machines trained on a synthesis-balanced dataset in classification of retinal OCT images

Outcomes of the classification of retinal OCT images trained on the synthesis-balanced dataset with pretrained DNNs are depicted in Table 4.

Table 4 Experimental results of deep learning machines trained on a synthesis-balanced dataset on selected pretrained DNNs in classification of retinal OCT images (experiment 2)

	Accuracy (%)	Sensitivity (%)	Specificity (%)	Precision (%)	F1 score (%)	MCC (%)
AlexNet						
Training set	98.14	98.14	99.38	98.15	98.14	97.52
Validation set	97.23	97.23	99.08	97.24	97.23	96.32
5-Fold CV	97.81	97.81	99.27	97.81	97.81	97.08
Kermany test set	99.38	99.38	99.79	99.39	99.38	99.18
TSGH test set	71.18	71.59	90.06	68.48	69.11	59.70
VGG19						
Training set	98.41	98.41	99.47	98.42	98.41	97.89
Validation set	98.07	98.07	99.36	98.07	98.07	97.43
5-Fold CV	97.99	97.99	99.33	97.99	97.99	97.32
Kermany test set	99.38	99.38	99.79	99.39	99.38	99.18
TSGH test set	80.45	76.64	92.62	80.62	78.17	71.36
ResNet101						
Training set	98.60	98.60	99.53	98.62	98.60	98.14
Validation set	98.41	98.41	99.47	98.43	98.41	97.89
5-Fold CV	98.52	98.52	99.51	98.53	98.52	98.03
Kermany test set	99.38	99.38	99.79	99.38	99.38	99.18
TSGH test set	84.92	84.52	94.43	83.81	84.01	78.63

CV: Cross-Validation

For the training set, ResNet101 was the top classifier with an accuracy of 98.60%, sensitivity of 98.60%, specificity of 99.53%, precision of 98.62%, F1 score of 98.14%, and MCC score of 98.13%. In terms of accuracy, it is more than 0.8% higher in comparison with the unbalanced dataset's top classifier for its training set.

In terms of the validation set, ResNet101 was the top classifier with an accuracy of 98.41%, sensitivity of 98.41%, specificity of 99.47%, precision of 98.43%, F1 score of 98.41%, and MCC score of 97.89%. An improvement of almost 2% in respect to accuracy compared with the unbalanced dataset's top classifier for its validation set.

With fivefold CV (cross validation), pretrained DNN Resnet101 was the top classifier, with an accuracy of 98.52%, sensitivity of 98.52%, specificity of 99.51%, precision of 98.53%, F1 score of 98.52%, and MCC of 98.03%; which presents a more than 1% betterment in terms of accuracy in comparison with the unbalanced dataset's top classifier for its training set.

With the Kermany test set, AlexNet and VGG19 were the top classifiers with congruent accuracies of 99.38%, sensitivities of 99.38%, specificities of 99.79%, precisions of 99.39%, F1 scores of 99.38%, and MCC scores of 99.18%.

When tested with the TSGH test set, ResNet101 was the top classifier with an accuracy of 84.92%, sensitivity of 84.52%, specificity of 94.43%, precision of 83.81%, F1 score of 84.01%, and MCC score of 78.63%. An improvement of over 3.5% for accuracy in comparison to the unbalanced dataset's top classifier for its TSGH test set.

The machine performance-enhancing capacity of GAN is evident as machines trained on the synthesis-balanced dataset outdid those trained on the unbalanced dataset with the training set, validation set, fivefold CV, and TSGH test in terms of classification accuracy.

Experiment #3: Performance of deep learning machines trained on the reduction-balanced dataset in classification of retinal OCT images

Outcomes of the classification of retinal OCT images trained on the reduction-balanced dataset with pretrained DNNs are depicted in Table 5.

For the training set, pretrained DNN ResNet101 was the top classifier, with an accuracy of 97.88%, sensitivity of 97.88%, specificity of 99.29%, precision of 97.93%, F1 score of 97.89%, and MCC score of 97.20%.

In terms of the validation set, pretrained DNN ResNet101 was the top retinal OCT image classifier, with an accuracy of 96.50%, sensitivity of 96.50%, specificity of 98.83%, precision of 96.59%, F1 score of 96.51%, and MCC score of 95.38%.

With fivefold CV (cross validation), pretrained DNN ResNet101 was the top classifier, with an accuracy of 97.44%, sensitivity of 97.44%, specificity of 99.15%, precision of 97.45%, F1 score of 97.44%, and MCC of 96.59%.

With the Kermany test set, pretrained DNN ResNet101 was the top classifier, with an accuracy of 99.69%, sensitivity of 99.69%, specificity of 99.90%, precision of 99.69%, F1 score of 99.69%, and MCC of 99.59%. An improvement of

Table 5 Experimental results of deep learning machines trained on a reduction-balanced dataset on selected pretrained DNNs in classification of retinal OCT images (experiment 3)

	Accuracy (%)	Sensitivity (%)	Specificity (%)	Precision (%)	F1 score (%)	MCC (%)
AlexNet						
Training set	95.13	95.13	98.38	95.50	95.13	93.67
Validation set	93.08	93.08	97.69	93.60	93.07	91.01
5-Fold CV	95.24	95.24	98.41	95.34	95.24	93.69
Kermany test set	98.86	98.86	99.62	98.87	98.86	98.49
TSGH test set	66.62	71.01	88.68	65.78	66.34	56.46
VGG19						
Training set	96.47	96.47	98.82	96.58	96.48	95.34
Validation set	95.31	95.31	98.44	95.51	95.33	93.84
5-Fold CV	95.48	95.48	98.49	95.56	95.49	94.00
Kermany test set	99.59	99.59	99.86	99.59	99.59	99.45
TSGH test set	77.81	75.76	91.57	77.34	74.76	67.81
ResNet101						
Training set	97.88	97.88	99.29	97.93	97.89	97.20
Validation set	96.50	96.50	98.83	96.59	96.51	95.38
5-Fold CV	97.44	97.44	99.15	97.45	97.44	96.59
Kermany test set	99.69	99.69	99.90	99.69	99.69	99.59
TSGH test set	78.77	72.57	91.72	82.17	73.34	68.32

CV cross-validation

0.10% in respect to accuracy in comparison with the unbalanced dataset's top classifier for its Kermany test set.

When tested with the TSGH test set, pretrained DNN ResNet101 was the top classifier, with an accuracy of 78.77%, sensitivity of 72.57%, specificity of 91.72%, precision of 82.17%, F1 score of 73.34%, and MCC score of 68.32%.

With the exception of the Kermany test set, machines trained on the synthesis-balanced dataset achieved higher accuracies than machine trained on reduction-balanced for all sets. Overall, the highest accuracies for all sets were either trained on the synthesis-balanced or reduction-balanced datasets.

Performance of ROC-AUC curves

ROC-AUC curves of deep learning machines trained on different datasets with selected pretrained DNNs were drawn for classes: CNV, DME, DRUSEN, and NORMAL for the TSGH test set and are shown in Fig. 4.

For the unbalanced dataset: on the AlexNet Network, AUCs of 0.8275, 0.8552, 0.8798, and 0.9117 were achieved for classes, CNV, DME, DRUSEN, and NORMAL, respectively. On the VGG19 network, AUCs of 0.8798, 0.9194, 0.9819, and 0.9690 were reached for classes, CNV, DME, DRUSEN, and NORMAL, respectively. On the ResNet101 network, AUCs of 0.9298, 0.9378, 0.9914, and 0.9759 were attained for classes, CNV, DME, DRUSEN, and NORMAL, respectively.

For the synthesis-balanced dataset: on the AlexNet Network, AUCs of 0.8160, 0.8752, 0.9495, and 0.9169 were achieved for classes, CNV, DME, DRUSEN, and NORMAL, respectively. On the VGG19 network, AUCs of 0.8694, 0.9272, 0.9938, and 0.9599 were reached for classes, CNV, DME, DRUSEN, and NORMAL, respectively. On the ResNet101 network, AUCs 0.9324, 0.9536, 0.9991, and 0.9768 were attained for classes, CNV, DME, DRUSEN, and NORMAL, respectively.

For the reduction-balanced dataset: on the AlexNet Network, AUCs of 0.7477, 0.8285, 0.9713, and 0.8984 were achieved for classes, CNV, DME, DRUSEN, and NORMAL, respectively. On the VGG19 network, AUCs of 0.8800, 0.9058, 0.9878, and 0.9536 were reached for classes, CNV, DME, DRUSEN, and NORMAL, respectively. On the ResNet101 network, AUCs 0.8334, 0.9158, 0.9766, and 0.9620 were attained for classes, CNV, DME, DRUSEN, and NORMAL, respectively.

Illustrated by Fig. 5, AUCs obtained by classes trained on the ResNet101 network of the synthesis-balanced dataset were all over 0.9300. Moreover, with the exception of classes, CNV and DME from the AlexNet network, the AUCs for all machines trained on the synthesis-balanced dataset were all in the range of 0.9100.

Discussion

The main difference between the original StyleGAN and StyleGAN2-ADA is the integration of adaptive discriminator augmentation mechanisms, which helps overcome the issue of overfitting in small datasets. As stated in the original StyleGAN2-ADA literature,

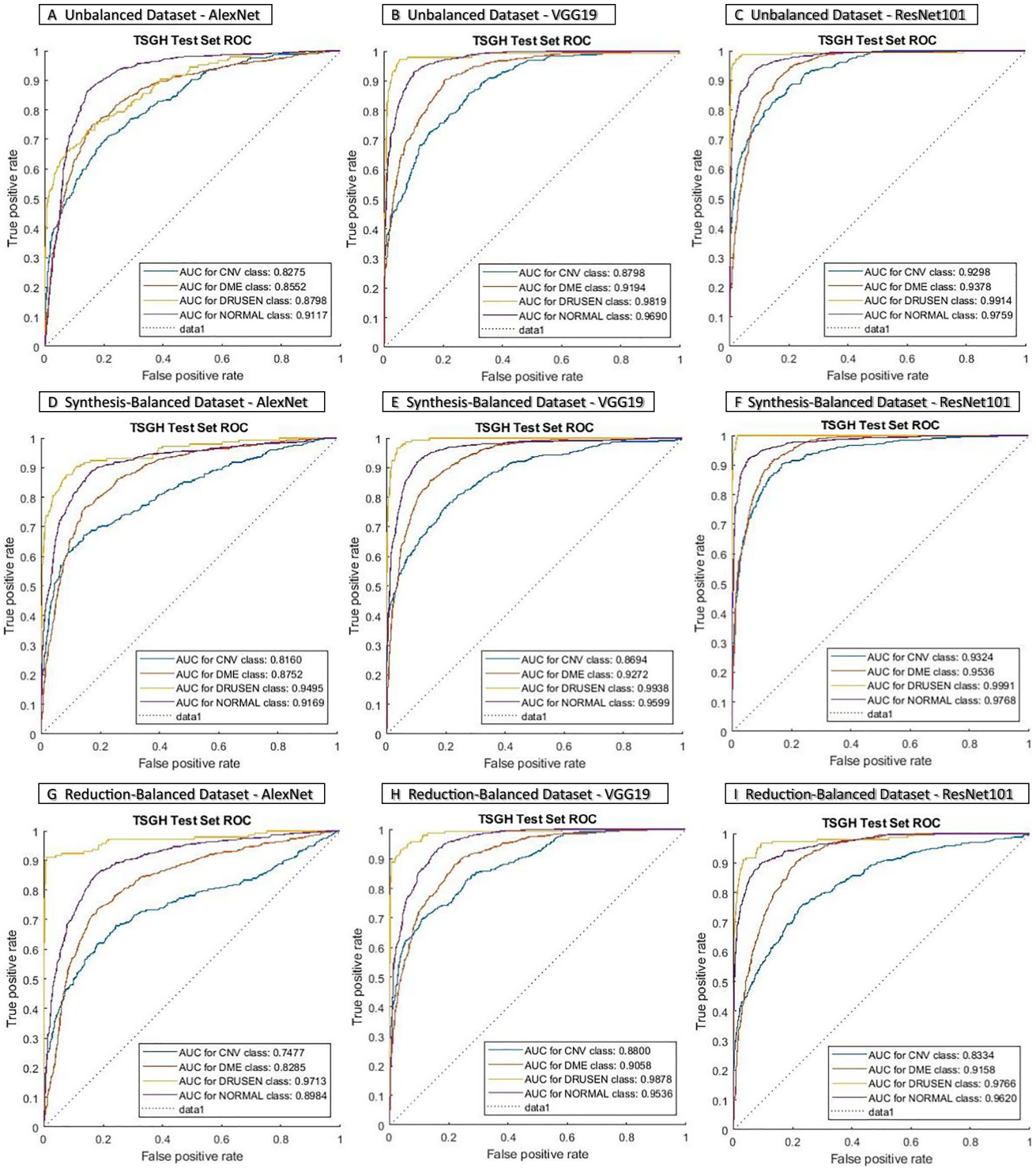
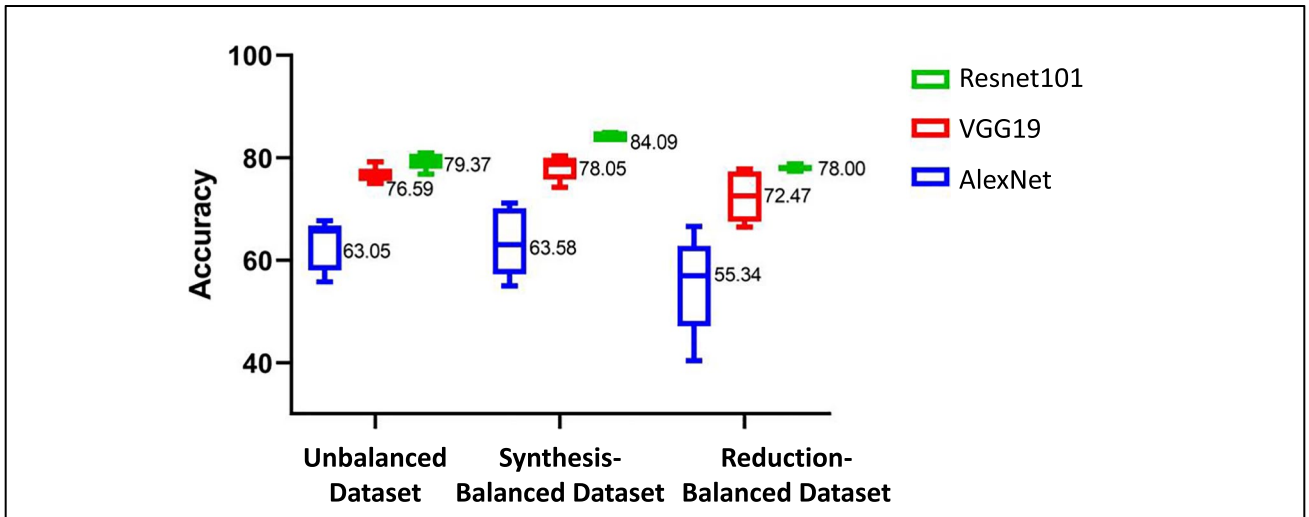


Fig. 4 ROC-AUC curves of deep learning machines trained on different datasets with selected pretrained DNNs in the classification of classes: CNV, DME, DRUSEN, and NORMAL for TSGH test set. **A**, **B**, **C** unbalanced dataset **D**, **E**, **F** synthesis-balanced dataset and **G**, H, **I** reduction-balanced dataset

differential discriminator augmentations for StyleGAN2-ADA impacts the final results [40]: from our experiment, augmentation "bgcfnc" seemed to be the most optimal configuration for conditions, DME, DRUSEN, and NORMAL. The literature also indicated transfer learning achieved better results than training from scratch, and



Pretrained DNN / Handle unbalanced	AlexNet						VGG19						ResNet101					
	fold1	fold2	fold3	fold4	fold5	mean	fold1	fold2	fold3	fold4	fold5	mean	fold1	fold2	fold3	fold4	fold5	mean
Unbalanced Dataset	67.72	55.76	65.78	65.71	60.28	63.05	79.20	76.51	76.27	75.02	75.94	76.59	79.44	76.80	80.64	81.03	78.96	79.37
Synthesis-Balanced Dataset	59.61	69.12	63.02	54.95	71.18	63.58	79.54	78.72	80.45	77.28	74.26	78.05	83.09	83.05	84.92	84.77	84.63	84.09
Reduction-Balanced Dataset	56.96	53.75	58.98	40.39	66.62	55.34	76.90	72.57	66.47	68.59	77.81	72.47	77.38	78.77	77.76	77.91	78.19	78.00

Fig. 5 5-Fold cross validation box and whisker plot and table of the performance of deep learning machines trained on different datasets with selected pretrained DNNs on the TSGH test set

thus was utilized in our experiment. In the preliminary assessment of fixed ordered discriminator augmentations step of the study, networks were trained to 8000 images because the smallest dataset consisted of 8000 images (rounded to the nearest available thousandths). In doing so, repetitive presentation of the same images to the discriminator is prevented and each picture is shown only once to the discriminator. This decision was made to ensure the networks of each class were assessed fairly.

The classification outcomes reached on the Kermany test set for machines trained on unbalanced, synthesis-balanced, and reduction-balanced datasets were greater than the outcomes achieved by their training set and validation set. After examining several related literatures, this phenomenon seemed to be present in other works as well [43–45] and is quite bizarre, as it defies a fundament of deep learning: classification outcomes on unseen images (Kermany test set) should not be greater than seen images (training set and validation set). In further investigation, two ophthalmologists examined the images of the Kermany test set and deemed the images to consist of prominent features for each of their respective classes, perhaps even more distinctive than images from the training set, hence the obscurity.

As depicted in Fig. 4 of the results section, in comparing the receiver operating characteristic (ROC) curves of machines' performance on the TSGH test set, the machine trained on the synthesis-balanced dataset with ResNet101 has distinctively larger areas for each class. Moreover, for all classes, curves were roughly analogous and areas were of similar size; thus, the machine presents an accurate and reliable capacity to classify the test set as a whole.

fivefold cross validation (CV) was integrated into our study to elevate the reliability of our results. This way, we can determine whether just one 80% and 20% split into training and validation yields favorable outcomes or a collective of differential splits can produce the same positive results. As portrayed in the Box and Whisker plot (Fig. 5) below, accuracies progressed from AlexNet (lowest) to VGG19 to ResNet101 (highest) for every dataset. Moreover, the whiskers of machines trained on the ResNet101 network do not extend far apart from the box for all datasets, and thus, the network is more reliable.

When deep learning machines trained on the unbalanced, synthesis-balanced, and reduction-balanced datasets were tested with the Kermany test set, which had the same origin as the training dataset, they were able to achieve high classification accuracies, uniform with similar studies of the same dataset [41, 43, 44, 46]. The classification accuracies achieved by our machines may not trump all relevant studies utilizing sets derived from Kermany's dataset. However, our study justifies the validity of balancing datasets—as all the highest accuracies achieved for all sets were by balanced datasets

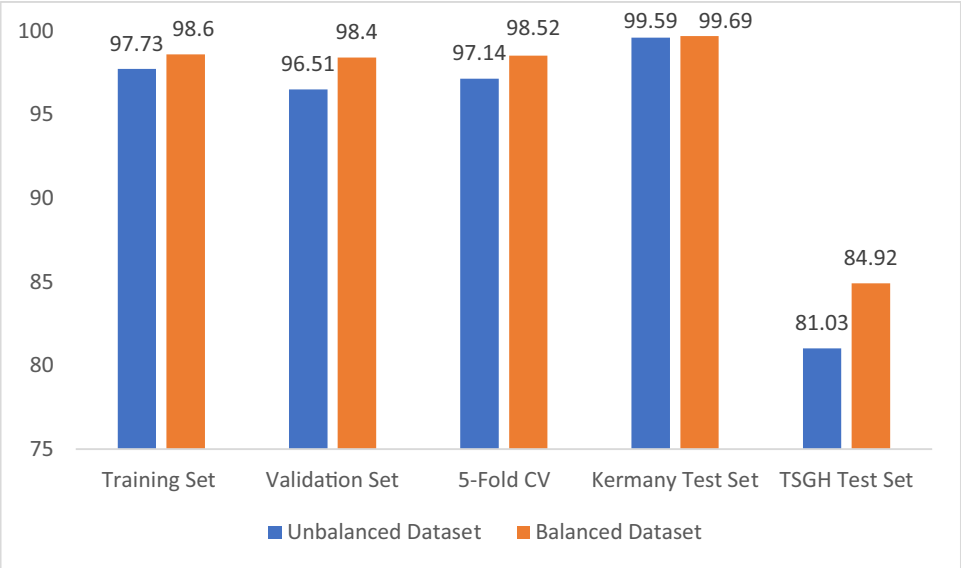


Fig. 6 Comparison of highest accuracies achieved with machines trained on unbalanced and balanced datasets

(synthesis-balanced and reduction-balanced)—and substantiates the feasibility of utilizing GANs to elevate classification accuracy, and applicable not only with sets derived from Kermany's dataset but also with a clinical set collected from a medical center in Taipei (shown in Fig. 6 below).

The number of DRUSEN images collected from TSGH was insignificant, this may be due to the often-asymptomatic nature of the condition; thereby patients do not seek medical attention and retinal OCT images do not end up being taken. The case of a low prevalence of DRUSEN in Chinese populations is seemingly unlikely, as a study in the UK (United Kingdom) of the prevalence of DRUSEN in Caucasians aged 18–54 years in gradable eyes is 91.48% [47]; however, population trends in one population are not always indicative of trends in another population. Ergo, a future study could further investigate DRUSEN trends in Chinese populations.

The scope of our current study only broadly scratches the surface of ophthalmology diagnoses; as each pathological lesion class can be further categorized into subclassifications. For example, DRUSEN can be subclassified into hard drusen, soft drusen, cuticular drusen, and subretinal drusenoid deposits. Hence, further studies could classify pathological retinal conditions not just into DME, DRUSEN, and CNV but into their distinctive subclassifications as well.

The training images of the Kermany dataset were labeled and verified through a 3-tier grading system of graders with increasing experience [41], and the final tier consisted of two senior retinal specialists with over 20 years of experience. Thus, retinal image pathologies should be true to their individual label. However, in clinical practice, there may be incidences of co-existing pathologies in one retinal image. To address such cases, in the future development of clinical OCT image diagnosis systems, on the basis there is no individual diagnosis with over 60% prediction percentage, machines could list the two most probable diagnoses and present them to the physician to make the final decision.

Our findings demonstrate the feasibility of utilizing GANs in balancing datasets to produce favorable classification outcomes. The technique—GAN—opens a pathway for the classification of disease datasets that do not possess an abundance of data; therefore, researchers may more deeply explore the use of deep learning in the classification of rare and thus poorly recorded diseases.

Conclusions

The utilization of a GAN to balance datasets used for deep learning training proves to be a viable methodology to enhance deep learning machines' performance. In the present day, diagnosis of a multitude of diseases heavily relies on imaging techniques, our study holds promise for future GAN-based deep learning approaches not only in ophthalmology but also in various clinical disciplines as well. In addition, the prospect of deep learning approaches reducing diagnostic burdens for ophthalmologists and elevating healthcare quality by primary physicians seems ever more hopeful.

Abbreviations *GAN*: Generative adversarial network; *OCT*: Optical coherence tomography; *CV*: Cross validation; *ADA*: Adaptive discriminator augmentation; *TSGH*: Tri-service General Hospital; *CNV*: Choroidal neovascularization; *DME*: Diabetic macular edema; *DNNs*: Deep neural networks; *FFHQ*: Flickr-Faces-High Quality; FFHQ512: Flickr-Faces-HQ (FFHQ) Dataset with resolution 512 × 512; *FID*: Frechet inception distance; *b*: Pixel blitting; *bg*: Pixel blitting, geometric transformation; *bgc*: Pixel blitting, geometric transformation, color transformation; *bgcf*: Pixel blitting, geometric transformation, color

transformation, image-space filtering; *bgcfn*: Pixel blitting, geometric transformation, color transformation, image-space filtering, additive noise; *bgcfnc*: Pixel blitting, geometric transformation, color transformation, image-space filtering, additive noise, cutout; *ADAM*: Adaptive moment estimation; *MCC*: Matthews correlation coefficient; *ROC*: Receiver operating characteristic curve; *AUC*: Area under the ROC curve; *UK*: United Kingdom

Acknowledgements Not applicable

Author contribution Drs. Chen and K.-F. Lin had full access to all of the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis. Concept and design: Sun, K.-F. Lin, Chen. Acquisition, analysis, or interpretation of data: Sun, Pao, Huang, Wei, K.-F. Lin, Chen. Drafting of the manuscript: Sun, Chen. Critical revision of the manuscript for important intellectual content: Sun, Pao, Huang, Wei, K.-F. Lin, Chen. Statistical analysis: Sun, Wei, K.-F. Lin, Chen. Obtained funding: Chen. Administrative, technical, or material support: Supervision: K.-F. Lin, Chen.

Funding The design and part-writing costs of the study are funded by the Ministry of Science and Technology, Taiwan (MOST 108–3111-Y-016–012) and costs of collection, analysis and interpretation of data and part-writing are funded by the Ministry of National Defense-Medical Affairs Bureau (MND-MAB-D-111097). Publication costs are funded by the authors.

Data availability The data that support the findings of this study are available from the corresponding author upon reasonable request.

Declarations

Ethics approval and consent to participate The experimental protocol was approved by the Tri-Service General Hospital (TSGH) human ethics committee under registration number IRB: 1–108-05–082. Images were retrospectively obtained from the Department of Ophthalmology and the Department of Endocrinology and Metabolism at TSGH and were anonymized; thus, informed consent was not required.

Consent for publication Not applicable.

Competing interests The authors declare no competing interests.

References

1. Bourne RR, Steinmetz JD, Saylan M et al (2021) Causes of blindness and vision impairment in 2020 and trends over 30 years, and prevalence of avoidable blindness in relation to VISION 2020: the right to sight: an analysis for the Global Burden of Disease Study. Lancet Glob Health 9:e144–e160. https://doi.org/10.1016/S2214-109X(20)30489-7
2. Rauch R, Weingessel B, Maca SM, Vecsei-Marlovits PV (2012) Time to first treatment: the significance of early treatment of exudative age-related macular degeneration. Retina 32:1260–1264. https://doi.org/10.1097/IAE.0b013e3182018df6
3. Ho AC, Albini TA, Brown DM et al (2017) The potential importance of detection of neovascular age-related macular degeneration when visual acuity is relatively good. JAMA Ophthalmology 135:268–273. https://doi.org/10.1001/jamaophthalmol.2016.5314
4. Good WV, Early Treatment for Retinopathy of Prematurity Cooperative Group (2004) Final results of the early treatment for retinopathy of prematurity (ETROP) randomized trial. Trans Am Ophthalmol Soc 102:233–248
5. Early Treatment For Retinopathy Of Prematurity Cooperative Group (2003) Revised indications for the treatment of retinopathy of prematurity: results of the early treatment for retinopathy of prematurity randomized trial. Arch Ophthalmol 121:1684–1694. https://doi.org/10.1001/archopht.121.12.1684
6. Neely DC, Bray KJ, Huisingh CE et al (2017) Prevalence of undiagnosed age-related macular degeneration in primary eye care. JAMA Ophthalmol 135:570–575. https://doi.org/10.1001/jamaophthalmol.2017.0830
7. Statham MO, Sharma A, Pane AR (2008) Misdiagnosis of acute eye diseases by primary health care providers: incidence and implications. Med J Aust 189:402–404. https://doi.org/10.5694/j.1326-5377.2008.tb02091.x
8. Yip H, Crock C, Chan E (2020) Diagnostic error in an ophthalmic emergency department. Diagnosis (Berl) 7:129–131. https://doi.org/10.1515/dx-2019-0047
9. Gelston CD, Patnaik JL (2019) Ophthalmology training and competency levels in care of patients with ophthalmic complaints in United States internal medicine, emergency medicine and family medicine residents. J Educ Eval Health Prof 16:25. https://doi.org/10.3352/jeehp.2019.16.25
10. Alotaibi AK, Alsalim A, Alruwaili F et al (2019) Burnout during ophthalmology residency training: a national survey in Saudi Arabia. Saudi J Ophthalmol 33:130–134
11. Cheung R, Yu B, Iordanous Y, Malvankar-Mehta MS (2021) The prevalence of occupational burnout among ophthalmologists: a systematic review and meta-analysis. Psychol Rep 124:2139–2154. https://doi.org/10.1177/0033294120954135
12. Rosenblatt TR, Vail D, Saroj N et al (2021) Increasing incidence and prevalence of common retinal diseases in retina practices across the United States. Ophthalmic Surg Lasers Imaging Retina 52:29–36. https://doi.org/10.3928/23258160-20201223-06
13. Li JQ, Welchowski T, Schmid M et al (2020) Prevalence and incidence of age-related macular degeneration in Europe: a systematic review and meta-analysis. Br J Ophthalmol 104:1077–1084. https://doi.org/10.1136/bjophthalmol-2019-314422
14. Ansah JP, Koh V, de Korne DF et al (2018) Projection of eye disease burden in Singapore. Ann Acad Med Singap 47:13–28
15. Lee R, Wong TY, Sabanayagam C (2015) Epidemiology of diabetic retinopathy, diabetic macular edema and related vision loss. Eye Vis (Lond) 2:17. https://doi.org/10.1186/s40662-015-0026-2
16. Jonas JB, Cheung CMG, Panda-Jonas S (2017) Updates on the epidemiology of age-related macular degeneration. Asia Pac J Ophthalmol (Phila) 6:493–497. https://doi.org/10.22608/APO.2017251
17. Medical Advisory Secretariat (2009) Optical coherence tomography for age-related macular degeneration and diabetic macular edema: an evidence-based analysis. Ont Health Technol Assess Ser 9:1–22
18. Fujimoto J, Swanson E (2016) The development, commercialization, and impact of optical coherence tomography. Invest Ophthalmol Vis Sci 57:OCT1-OCT13. https://doi.org/10.1167/iovs.16-19963
19. Arevalo JF, Lasave AF, Arias JD et al (2013) Clinical applications of optical coherence tomography in the posterior pole: the 2011 José Manuel Espino Lecture - Part I. Clin Ophthalmol 7:2165–2179. https://doi.org/10.2147/OPTH.S51098
20. Lee CS, Baughman DM, Lee AY (2017) Deep learning is effective for the classification of OCT images of normal versus age-related macular degeneration. Ophthalmol Retina 1:322–327. https://doi.org/10.1016/j.oret.2016.12.009
21. Li F, Chen H, Liu Z et al (2019) Fully automated detection of retinal disorders by image-based deep learning. Graefes Arch Clin Exp Ophthalmol 257:495–505. https://doi.org/10.1007/s00417-018-04224-8
22. Motozawa N, An G, Takagi S et al (2019) Optical coherence tomography-based deep-learning models for classifying normal

and age-related macular degeneration and exudative and non-exudative age-related macular degeneration changes. Ophthalmol Ther 8:527–539. https://doi.org/10.1007/s40123-019-00207-y
23. Schlegl T, Waldstein SM, Bogunovic H et al (2018) Fully automated detection and quantification of macular fluid in OCT using deep learning. Ophthalmology 125:549–558. https://doi.org/10.1016/j.ophtha.2017.10.031
24. Wu Q, Zhang B, Hu Y et al (2021) Detection of morphologic patterns of diabetic macular edema using a deep learning approach based on optical coherence tomography images. Retina 41:1110–1117. https://doi.org/10.1097/IAE.0000000000002992
25. Virgili G, Menchini F, Murro V, etal., (2011) Optical coherence tomography (OCT) for detection of macular oedema in patients with diabetic retinopathy. Cochrane Database Syst Rev 7:CD008081. https://doi.org/10.1002/14651858.CD008081.pub2
26. Lu W, Tong Y, Yu Y et al (2018) Deep learning-based automated classification of multi- categorical abnormalities from optical coherence tomography images. Transl Vis Sci Technol 7:41. https://doi.org/10.1167/tvst.7.6.41
27. Yoon J, Han J, Park JI et al (2020) Optical coherence tomography-based deep-learning model for detecting central serous chorioretinopathy. Sci Rep 10:18852. https://doi.org/10.1038/s41598-020-75816-w
28. Buda M, Maki A, Mazurowski MA (2018) A systematic study of the class imbalance problem in convolutional neural networks. Neural Netw 106:249–259. https://doi.org/10.1016/j.neunet.2018.07.011
29. Krawczyk B (2016) Learning from imbalanced data: open challenges and future directions. Prog Artif Intell 5:221–232. https://doi.org/10.1007/s13748-016-0094-0
30. He H, Garcia EA (2009) Learning from imbalanced data. IEEE Trans Knowl Data Eng 21(9):1263–1284. https://doi.org/10.1109/TKDE.2008.239
31. Kuwayama S, Ayatsuka Y, Yanagisono D et al (2019) Automated detection of macular diseases by optical coherence tomography and artificial intelligence machine learning of optical coherence tomography images. J Ophthalmol 2019:6319581. https://doi.org/10.1155/2019/6319581
32. Yoo TK, Choi JY, Seo JG et al (2019) The possibility of the combination of OCT and fundus images for improving the diagnostic accuracy of deep learning for age-related macular degeneration: a preliminary experiment. Med Biol Eng Comput 57:677–687. https://doi.org/10.1007/s11517-018-1915-z
33. Thakoor KA, Li X, Tsamis E et al (2021) Strategies to improve convolutional neural network generalizability and reference standards for glaucoma detection from OCT scans. Transl Vis Sci Technol 10:16. https://doi.org/10.1167/tvst.10.4.16
34. Choi KJ, Choi JE, Roh HC et al (2021) Deep learning models for screening of high myopia using optical coherence tomography. Sci Rep 11:21663. https://doi.org/10.1038/s41598-021-00622-x
35. Elgendi M, Nasir MU, Tang Q et al (2021) The effectiveness of image augmentation in deep learning networks for detecting COVID-19: a geometric transformation perspective. Front Med (Lausanne) 8:629134. https://doi.org/10.3389/fmed.2021.629134
36. Zheng R, Liu L, Zhang S et al (2018) Detection of exudates in fundus photographs with imbalanced learning using conditional generative adversarial network. Biomed Opt Express 9:4863–4878. https://doi.org/10.1364/BOE.9.004863
37. Guan S, Loew M (2019) Breast cancer detection using synthetic mammograms from generative adversarial networks in convolutional neural networks. J Med Imaging (Bellingham) 6:031411. https://doi.org/10.1117/1.JMI.6.3.031411
38. Xiao Y, Wu J, Lin Z (2021) Cancer diagnosis using generative adversarial networks based on deep learning from imbalanced data. Comput Biol Med 135:104540. https://doi.org/10.1016/j.compbiomed.2021.104540
39. Zheng C, Xie X, Zhou K et al (2020) Assessment of generative adversarial networks model for synthetic optical coherence tomography images of retinal disorders. Transl Vis Sci Technol 9:29. https://doi.org/10.1167/tvst.9.2.29
40. Karras T, Aittala M, Hellsten J, etal., (2020) Training generative adversarial networks with limited data. Adv Neural Inf Process Syst 33:1–15. https://doi.org/10.48550/arXiv.2006.06676
41. Kermany DS, Goldbaum M, Cai W et al (2018) Identifying medical diagnoses and treatable diseases by image-based deep learning. Cell 172:1122-1131.e9. https://doi.org/10.1016/j.cell.2018.02.010
42. The MathWorks, Inc (2022) Pretrained Deep Neural Networks - MATLAB & Simulink. Help Center https://www.mathworks.com/help/deeplearning/ug/pretrained-convolutional-neural-networks.html Accessed 2 May 2022
43. Tsuji T, Hirose Y, Fujimori K et al (2020) Classification of optical coherence tomography images using a capsule network. BMC Ophthalmol 20:114. https://doi.org/10.1186/s12886-020-01382-4
44. Chen YM, Huang WT, Ho WH, Tsai JT (2021) Classification of age-related macular degeneration using convolutional-neural-network-based transfer learning. BMC Bioinformatics 22:99. https://doi.org/10.1186/s12859-021-04001-1
45. Ho WH, Huang TH, Yang PY et al (2021) Artificial intelligence classification model for macular degeneration images: a robust optimization framework for residual neural networks. BMC Bioinformatics 22:148. https://doi.org/10.1186/s12859-021-04085-9
46. Sunija A, Kar S, Gayathri S et al (2021) Octnet: a lightweight cnn for retinal disease classification from optical coherence tomography images. Comput Methods Programs Biomed 200:105877. https://doi.org/10.1016/j.cmpb.2020.105877
47. Silvestri G, Williams MA, McAuley C et al (2012) DRUSEN prevalence and pigmentary changes in Caucasians aged 18–54 years. Eye (Lond) 26:1357–1362. https://doi.org/10.1038/eye.2012.165