IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received June 28, 2021, accepted July 3, 2021, date of publication July 7, 2021, date of current version July 15, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3095391

Studying the Applicability of Generative Adversarial Networks on HEp-2 Cell Image Augmentation

ASAAD ANAAM, HANI M. BU-OMER, (Graduate Student Member, IEEE), AND AKIO GOFUKU

Graduate School of Interdisciplinary Science and Engineering in Health Systems, Okayama University, Okayama 700-8530, Japan

Corresponding author: Asaad Anaam (asaadanam@s.okayama-u.ac.jp)

ABSTRACT The Anti-Nuclear Antibodies (ANAs) testing is the primary serological diagnosis screening test for autoimmune diseases. ANAs testing is conducted mainly by the Indirect Immunofluorescence (IIF) on Human Epithelial cell-substrate (HEp-2) protocol. However, due to its high variability, human-subjectivity, and low throughput, there is an insistent need to develop an efficient Computer-Aided Diagnosis system (CADs) to automate this protocol. Many recently proposed Convolutional Neural Networks (CNNs) demonstrated promising results in HEp-2 cell image classification, which is the main task of the HE-p2 IIF protocol. However, the lack of large labeled datasets is still the main challenge in this field. This work provides a detailed study of the applicability of using generative adversarial networks (GANs) algorithms as an augmentation method. Different types of GANs were employed to synthesize HEp-2 cell images to address the data scarcity problem. For systematic comparison, empirical quantitative metrics were implemented to evaluate different GAN models' performance of learning the real data representations. The results of this work showed that though the high visual similarity with the real images, GANs' capacity to generate diverse data is still limited. This deficiency in the generated data diversity is found to be of a crucial impact when used as a standalone method for augmentation. However, combining limited-size GANs-generated data with classic augmentation improves the classification accuracy across different variants of CNNs. Our results demonstrated a competitive performance for the overall classification accuracy and the mean class accuracy of the HEp-2 cell image classification task.

INDEX TERMS Computer-aided diagnosis systems (CADs), convolutional neural networks (CNNs), data augmentation, data diversity, evaluation metrics, generative adversarial networks (GANs), HEp-2 cell image classification.

I. INTRODUCTION

Antinuclear autoantibodies (ANAs) testing plays a pivotal role in the serological diagnosis of autoimmune diseases [1] such as Systemic Lupus Erythematosus, Sjorgren's syndrome, and Rheumatoid Arthritis, etc. In this respect, Indirect Immunofluorescence (IIF) using the human Epithelium larynx carcinoma substrate (HEp-2) is considered the ''gold-standard'' protocol for ANAs testing [2]. However, this protocol of classifying the staining patterns of the HEp-2 cells suffers from high variability, observer-subjectivity, and low throughput. Therefore, there was an insistent need for developing an efficient Computer-Aided Diagnosis system (CAD) system to overcome these issues [1].

In this regard, a considerable amount of research was introduced to develop efficient HEp-2 cell pattern classifiers, mainly in the HEp-2 cell classification contests [3]–[6]. While the early works tried to handle this task based on hand-crafted feature methods, the recently proposed CNNs-based classifiers demonstrated superiority in HEp-2 cell image classification [7]. However, a large amount of labeled data is required for efficiently training Convolutional Neural Networks (CNNs) to avoid overfitting and improve the generalization capability, which is a challenging problem in many medical imaging fields. Collecting accurately annotated data is a complex and resources-consuming task

The associate editor coordinating the review of this manuscript and approving it for publication was Yizhang Jiang.

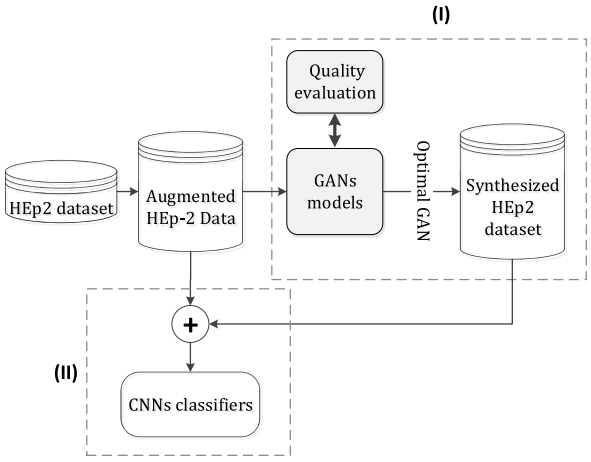


FIGURE 1. Block diagram of the proposed approach for data augmentation using GANs. Phase I: Different variants of GANs are trained and evaluated. Phase II: The best performing GAN is used for augmentation for CNNs classifiers.

for medical applications. Thus, various data augmentation methods are commonly used to enlarge the data size and alleviate this problem. Recently, since the unfolding of the generative adversarial networks (GANs) [8], it gained increasing research attention in the medical imaging fields, including image segmentation, registration, translation, synthesis, and classification [9]. Mainly, the continuously improving capabilities of GANs algorithms to approximate the real data distributions are of high interest to the recent studies in the medical imaging fields to address the problem of size limitations of annotated medical images.

This paper provides an investigation study of the effectiveness of using GANs approaches for synthesizing HEp-2 cell images for data augmentation purposes. We adopted two cascaded sets of experiments as described in the block diagram depicted in Fig. 1. In Phase I, four different approaches of GANs, characterized by their well-stability, were implemented to generate HEp-2 cell images, particularly: Deep Convolutional GAN (DCGAN) [10], Wasserstein GAN (WGAN) [11], Wasserstein GAN with gradient penalty (WGANGP) [12], And a modified version of WGANGP (annotated here as Info-WGANGP) [13] that inherits the mutual information maximization loss function of the Info-GAN [14]. To evaluate the quality of the generated images, two robust empirical quantitative metrics were used, which are the Fréchet Inception Distance (FID) [15] and 1-Nearest Neighbor classifier in two-sample tests [16]. These metrics provide the capability to analyze the statistical characteristics of the generated and real images and conduct a systematic assessment and comparison between the different implemented GANs models. Phase II includes experiments to compare the classification performances of some of the HEp-2 cell classification state-of-the-art CNN models [17]–[20] and the Inception-V3 model [21] when different data augmentation methods are used for training. This set of experiments evaluated the effectiveness of augmenting HEp-2 cell data with optimal GAN-synthesized images.

The contribution of this work could be summarized as:

1) We implemented four different GANs variants, characterized by their training stability, for the task of learning the HEp-2 cell images' visual representation. This aims to explore the learning capacity and the training robustness of the different GANs configurations in imitating the visual representations of such real-world data.
2) We applied two robust empirical GANs evaluation metrics to quantitatively assess the performances of the implemented GANs and evaluate their capabilities to approximate the real data distribution. These quantitative metrics provide systematic tools to understand the capabilities and limitations of the GANs variants under study.
3) We investigated the applicability of using GANs-generated images for augmentation by evaluating the performance of various HEp-2 cell images state-of-the-art CNNs classifiers [17]–[20] and the Inception-V3 model [21] trained with different augmentation methods. Multiple CNNs classifiers with different architectures were implemented to particularly inspect the effectiveness of the GANs-generated images regardless the CNNs configuration.

In the next section, the related work in HEp-2 cell image classification and GANs for medical image synthesis is introduced. Then, a brief description of the proposed GANs models and the evaluation metrics were introduced in the methods section. The experiments section describes the dataset and preprocessing, the classic augmentation methods, and the experimental details of each phase. Finally, the achieved results are demonstrated and discussed in section V and concluded in section VI.

II. RELATED WORK

A. CNNs FOR HEp-2 CELL IMAGE CLASSIFICATION

Unlike the conventional machine learning approaches, CNNs-based methods have the advantage of offering an automatic feature-learning process that demonstrates superiority over the hand-crafted ones for the HEp-2 cell classification task. One of the earliest works on this topic was proposed by Gao *et al.* [22] using a shallow CNNs model. Despite its unpowerful results, their study revealed interesting observations about the important rule of rotation in HEp-2 cell image augmentation and the informative nature of their extracellular textures for classification learning. Bayramoglu *et al.* [23] used AlexNet architecture with various approaches of pre-processing and data augmentation, and Jia *et al.* [24] adopted a customized CNN model that shares the general structure of the VGG network [25] configuration.

A deeper CNN model called the Deep Residual Inception Network (DRI-Net) was proposed by Li and Shen [18] that merges the general architectural configuration of both

Inception-net [21] and ResNet [26]. DRI-net integrates the multi-scale feature extraction property of Inception-net and the efficient network optimization advantage of ResNet. In other prominent work, Lei *et al.* [20] introduced a pre-trained ResNet50-based model called Deeply Supervised Residual Networks (DSRNet) that combines three depth-stages layer predictions into the final classification layer of the ResNet50 architecture and applies a transfer learning among two different datasets. This method represents one of the state-of-the-art performances for the HEp-2 cell classification task at the cost of significantly increasing the number of learnable parameters.

Another state-of-the-art result was achieved by a customized residual-based CNN model called Deep-Cross Residual Network (DCR-Net) with an intensive data augmentation approach suggested by Shen *et al.* [17]. In a recent work, Yuexiang and Linlin [19] introduced a fully customized lighter-weight network called ''HEpNet'' specifically to solve this problem. HEpNet is built from a small module called multi-scale convolutional component (MCC) which composed of different scales dilated convolutional layers and one shortcut connection. HEpNet demonstrated a high capability of learning representative features from fewer HEp-2 data, achieving competitive performance with less augmented data and shorter training time. Recently, Vununu *et al.* [27] proposed the use of the Discrete Wavelet Transform (DWT) as a pre-processing stage to capture more discriminative information than that of the spatial space of the original images. In their method, four DWT coefficient images are obtained for each original image before feeding them into a parallel-stream network paradigm that achieved competitive results.

While the studies mentioned above adopted the end-to-end configurations, other works proposed using CNN as a feature extractor, followed by a separated classifier. For example, Lu *et al.* [28] used the VGG16 network to learn the representative features and SVM-RBF for classification. Vununu *et al.* [29] suggested using two distinct VGG-like convolutional autoencoders (CAE) to extract two levels of features from regular and gradient images. The two features are then combined and classified using a simple neural network-based classifier. Recently, Cascio *et al.* [30] adopted a two-phases classifier framework for features extracted based on a pre-trained AlexNet network. A comprehensive review in this topic could be found in [7].

B. GANs FOR MEDICAL IMAGE CLASSIFICATION

Many recent studies in the field of medical image classification proposed using GANs to address scarcity in the annotated training data. In this regard, two general groups of studies using GANs in medical image classification tasks could be recognized. The first group is related to the works that adopted GANs for synthesizing new images from the real data domain and then train CNNs classifiers as a separated step. For example, a study for synthesizing artificial chest X-ray images to balance and augment modest size real datasets using DCGAN showed improvement in the classifier performance [31]. In another work, Frid-Adar *et al.* [32] also implement class-wise DCGAN models to enlarge the data size of liver lesions images which improved the classification performance when concatenating with the real training data. Baur *et al.* [33] studied using a modified version of LAPGAN to generate skin lesions images. Moreover, using CGAN was proposed by Finlayson *et al.* [34] for detecting bone fractures from pelvic radiographs.

The works in the second group adopted GANs in a semi-supervised manner. In these approaches, the trained GANs' discriminators are used as independent classifiers. For example, Hu *et al.* [13] proposed training a combined WGANGP [12] and Info-GAN [14] framework in an unsupervised manner for cell-level feature representation learning in the histopathology images classification task. Lecouat *et al.* [35] applied a patch-based semi-supervised GAN learning approach to classify diabetic retinopathy from funduscopic images. In the field of cardiac abnormality classification using X-ray imaging, Madani *et al.* [36] proposed a GAN-based semi-supervised learning approach to improve the classification performance with less annotated data. Moreover, in cellular structure image classification, Wang *et al.* [37] transferred a trained GAN's discriminator network into a new Alex-style CNN before fine-tuning with real images for improving the classification performance. More studies on this field are summarized in these reviews [9], [38].

C. GANs FOR HEp-2 IMAGE CLASSIFICATION

Using GANs for addressing HEp-2 image tasks is still in its preliminary stage, as just a few works were published in this area. One of the early studies used GANs for HE-p2 cell image segmentation was proposed by Li and Shen [39]. They used a U-net generator based on a modified framework that combines both pix2pix [40] and ACGAN [41] with a transfer-learning technique to boost the segmentation performance across different cell modalities. In the field of HEp-2 cell image synthesis, Kastaniotis *et al.* [42] proposed using a Teacher-network to guide the attention maps in the discriminator hidden layers in DCGAN [10] framework to improve the quality of the generated HEp-2 cell images. However, no evaluation measures have been applied in their study. In a recent work, Gupta *et al.* [43] applied the DCGAN framework with slight modification in the models' architectures to generate synthetic samples of the minor HEp-2 mitotic class. As demonstrated in their work, augmenting the minor mitotic class with GAN-synthesized images shows promising results to alleviate the problem of unbalanced data of the HEp-2 mitotic/interphase classification task. Furthermore, Xie *et al.* [44] used pix2pix [40] like GAN model to generate mask images of the Hep-2 cell-level images as a pre-stage before inputting pairs of the original images and their corresponding generated masks into a modified ResNet-50 classifier. Adding the generated mask images are suggested to enrich the classification network with more

boundary information. However, their results showed less classification performance comparing to the current state-of-the-art CNN approaches.

To the best of our knowledge, the work proposed by Majtner *et al.* [45] is the only published study that has proposed to explore using GAN-based synthesized images as a data augmentation method for HEp-2 cell-level classification task. In their work, an individual DCGAN [10] model was trained for each HEp-2 class in the I3A dataset [4] to cope with the high within-class heterogeneity of HEp-2 data. For evaluation, performance comparison has been conducted using three generic CNN configurations trained with different data augmentation methods. Their results revealed that DCGAN-based synthesized images were found to be less effective than those obtained by classic augmentation routines for training CNNs classifiers. However, the mentioned study investigated only the use of DCGAN [10], which does not yield the best results in many real-world data generating tasks [13]. Furthermore, no quantitative evaluation metrics were applied to assess the quality of the generated images and understand to which extent GANs could successfully approximate the underlying distribution of the real data.

To fill this gap, this paper aims to extend the research in this area and provide a deeper study on the capabilities and limitations of applying different GANs variants for HEp-2 cell images augmentation purposes.

III. METHODS

A. GANs FOR HEp-2 CELL IMAGES SYNTHESIS

Generative adversarial networks (GANs) [8] are recently proposed generative models that are used to learn the data representation in an unsupervised manner. GANs are composed of two competing CNNs. The first is the generator (G) which is a scale-up CNNs that is designed to transform a noise vector z (sampled from known distribution p_z) into an image space data $G(z)$ that is similar to the real images. The second is the discriminator (D) which is a binary classifier network trained to correctly distinguish between the real and the generated images ($D(.)$ outputs a probability of 1 for real input and 0 for generated images). The two competing models are trained together in an adversarial zero-sum game until the convergence takes place when G can generate plausible examples and D becomes just 50% certain about the input image source. Many GANs variants were proposed pursuing to ensure converging training and efficient representation-learning capacity. In this work, we examined four well-stable variants of GANs in the task of learning the visual representation of HEp-2 cell images, which are:

1) DCGAN

The original architecture of Deep Convolutional GAN (DCGAN) [10] was trained by optimizing the original GAN loss function of a two-player minimax game:

$$\min_G \max_D \left[E_{x\sim p_r} \log\left(D(x)\right) + E_{z\sim p_z} \log\left(1 - D(G(z))\right)\right], \quad (1)$$

where x is a real image sampled from unknown real data distribution p_r, z is a latent noise vector sampled from a noise distribution p_z, and $G(z)$ is the generated image.

2) WGAN

Wasserstein GANs [11] uses Earth Mover distance between the real data distribution and the generator distribution $W(p_r, p_g)$ to formulate the WGAN objective function. It turns out that the D (called critic) is trained to maximize this distance while G is trained to minimize it. The objective function of WGAN is written as follows:

$$\min_G \max_{D\in\mathcal{D}} \left[E_{x\sim p_r}\left[D(x)\right] + E_{z\sim p_z}\left[D(G(z))\right]\right], \quad (2)$$

where $\mathcal{D}$ is the set of 1-Lipschitz functions. However, to maintain the Lipschitz constraint, the weight of the critic D is explicitly clipped within a compact space $[-l, l]$. While WGAN demonstrated good training stability, it could not be guaranteed to converge with very deep architectures. Thus, the WAGN objective function was implemented using the DCGAN [10] model architecture.

3) WGANGP

Wasserstein GAN with gradient penalty [12] was proposed to improve WGAN by introducing a gradient penalty on the discriminator to ensure maintaining the continuity condition within the space of 1-Lipschitz functions. The objective function of the WGANGP is written as:

$$\begin{aligned}\min_G \max_{D\in\mathcal{D}} &\left[E_{x\sim p_r}\left[D(x)\right] + E_{z\sim p_z}\left[D(G(z))\right]\right] \\ &+\lambda E_{x\sim p_r, z\sim p_z, \alpha\sim(0,1)}[(\|\nabla D(\alpha x + (1-\alpha)G(z))\|_2 - 1)^2],\end{aligned} \quad (3)$$

where λ is a regularization hyperparameter, α is an interpolation vector of values between 0 and 1. ResNet-based network architecture as proposed in [12] was used to implement this model.

4) INFO-WGANGP

As proposed in [13], the WGANGP [12] objective function is unified with the information maximization functionality proposed in Info-GAN [14] to form a new hybrid GAN loss that takes advantage of both models. Introducing mutual information condition into the stable WGANGP loss formulation enforces the model to learn interpretable and disentangled representations underlying the data distribution in correspondence to a chosen latent variable c. For optimizing this objective function, an auxiliary classifier network Q is used to maximize the mutual information between the latent random variable c and the visual features of the generated samples $G(z, c)$. Thus, the objective function is defined as follow:

$$\begin{aligned}\min_{G,Q} \max_{D\in\mathcal{D}} &E_{x\sim p_r}[D(x)] \; + \; E_{z\sim p_z}[D(G(z))] \\ &+ \lambda_1 \, E_{x\sim p_r, z\sim p_z, \alpha\sim(0,1)}[(\|\nabla D(\alpha x + (1-\alpha)G(z))\|_2 - 1)^2] \\ &- \lambda_2 \, E_{z\sim p_z, c\sim p_c}[\log Q(c|G(z, c))],\end{aligned} \quad (4)$$

where λ_1 and λ_2 are hyperparameters, and c is a categorical latent variable sampled from a fixed noise distribution $p(c)$. Info-WGANGP model was implemented using ResNet-based network as proposed in [13]. In this model, the discriminator (critic) architecture was extended to have two output layers. The first is of a single dimension corresponding to the Wasserstein distance ($D(.)$), and the other is of c dimensions representing the Q network that predicts the category of the input images.

B. GANs EVALUATION METRICS

Seeking quantitative interpretable measures for ranking and analyzing the GANs performance, we implemented two samples-based metrics that had demonstrated their robustness in the literature of GANs evaluation, as follows:

1) FRÉCHET INCEPTION DISTANCE (FID)

FID [15] measures the similarity between two sample sets based on their Fréchet distance in an embedded space. The embedding is computed using the Inception V3 network [21] fixed up to a specific layer. The distributions of both real and generated images are assumed to follow a multivariate normal distribution which is estimated by computing their means and covariances. In particular, the FID is computed as:

$$FID(\chi_1, \chi_2) = \|\mu_1 - \mu_2\|_2^2 + Tr(\Sigma_1 + \Sigma_2 - 2(\Sigma_1\Sigma_2)^{\frac{1}{2}})), \tag{5}$$

where χ_1 and χ_2 are the embedding features of the real and generated sets, respectively. The μ_1 *and* Σ_1 refer to the mean vector and the covariance matrix of the real set features, respectively. Likewise for the generated set features, μ_2 *and* Σ_2 are the mean vector and the covariance matrix. Since FID scores measures distance, the lower the FID scores, the better the generated images are.

2) 1-NEAREST NEIGHBOR CLASSIFIER IN TWO-SAMPLE TEST

1-NN classifier in two-sample test [16] is a particular type of classifier two-sample test (C2ST) family [46], which are used to statistically assess whether two sample sets belong to the same distribution. This metric provides a useful tool to understand how the generator and the real distributions differ in interpretable units. The 1-NN classifier in two-sample test was implemented as follows: 1) Two equally sized sets S_r and S_g were sampled from a holdout real data distribution p_r and the generator distribution p_g, respectively. 2) The embedding space of both sets is computed using a pre-trained HEpNet [19] model to linearize the image manifolds. 3) A binary dataset is constructed by assigning positive labels to the real samples and negative labels to the generated samples. 4) The leave-one-out (LOO) accuracy of a binary 1-NN classifier is computed. According to this setup, the 1-NN classification accuracy is interpreted as follows:

- When both sample sets are drawn from identical data distributions, the 1-NN accuracy should remain around 50% (the chance-level accuracy). This is the ideal scenario for GANs, indicating that p_g is perfectly matching p_r.
- Straightforward, as the 1-NN accuracy increases over 50%, that unveils increasing in the distribution differences between p_r and p_g. However, 1-NN accuracy lower than 50% indicates that GAN overfits p_g to S_r.

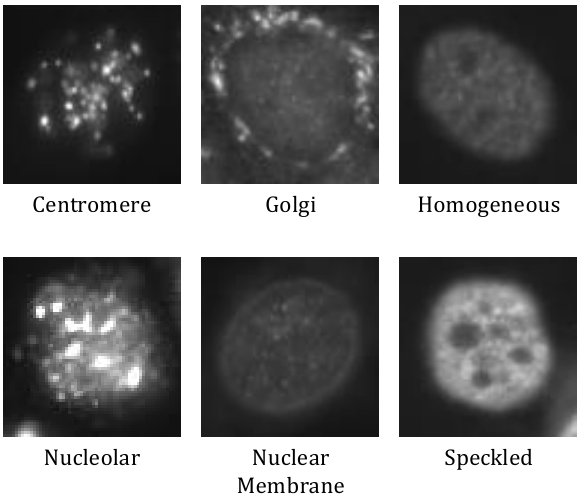


FIGURE 2. Examples of images from the I3A dataset showing different classes of HEp-2 cells.

However, more information about the properties of the generator distribution could be achieved when analyzing the LOO accuracy for each class separately. Particularly, for the real-class LOO accuracy (true positive rate), if it is relatively high, that might indicate that GAN suffers from a mode dropping because the generative model does not capture some portions (modes) of the real data distribution. However, if it is relatively low (around 50%), that would be because GANs-generated samples could cover all modes in the real distribution, which is the desired scenario. On the other hand, examining the generated-class accuracy (true negative rate) reveals information about the mode collapsing. In the mode collapse scenario, this measure is expected to be high as the generated samples tend to cluster in a few centers, increasing the generated-class accuracy. These three measures were annotated in the evaluation experiments of this study as 1-NN accuracy, 1-NN real-class accuracy, and 1-NN generated-class accuracy.

IV. EXPERIMENTS

As demonstrated in the block diagram shown in Fig. 1, a paradigm of two cascaded phases was adopted for using GAN as an augmentation method. The first phase involved training GANs to learn the visual representations of HEp-2 cell images and evaluate their performance. In the second phase, the effectiveness of using the optimal GAN model for augmentation was examined using some of the state-of-the-art CNNs. In this section, we introduced the used dataset, preprocessing, the classic augmentation method, and then discussed the experimental details of each phase.

A. DATASET AND PREPROCESSING

This study was conducted using the I3A dataset,[1] that was introduced in ICIP2013 [4] and then subsequently reused in ICPR2014 [5], and ICPR2016 [6]. While the test portion of this dataset is kept private for evaluation purpose, the training set containing 13,596 monochrome pre-extracted and annotated cell images were made publicly available. This dataset consists of six classes: Centromere, Golgi, Homogeneous, Nucleolar, Nuclear Membrane, and Speckled. Hereafter annotated as Ce, Gl, Ho, Nu, NuM, and Sp, respectively. Fig. 2 shows an example image from each class. The dataset was randomly partitioned into 64%, 16%, and 20% for training, validation, and testing, respectively. The training partition is used for training networks in both phases' experiments: training GANs models, and training CNNs classifiers. As well, the validation set is also used for evaluating the training performance in the experiments of both phases. The test set is kept for evaluating the final CNNs classifiers' performance. Table 1 summarizes the number of images in each partition per class. The HEp-2 IIF images demonstrated high variance in the image intensity, which correlates to the strength of the pattern of each particular sample [1]. To alleviate the intensity variation severity of the HEp-2 cell images, a contrast stretching was applied using a method represented by the following equation:

$$I_o = \frac{I_{in} - c}{d - c} \times 255, \tag{6}$$

where I_o is the output image, I_{in} is the input images, c and d are pixels values of the 1^{st} and the 99^{th} percentile of the I_{in} histogram, respectively. Furthermore, all images were resized to the dimension of 64 $\times$ 64 pixels using bicubic interpolation.

B. CLASSIC AUGMENTATION

An intense classic augmentation method similar to that in [17] was implemented for two reasons. The first is to train both the GANs and the CNNs classifiers efficiently. The second is to examine the effectiveness of using GANs-generated images for training CNNs compared to the highest limit of the classic augmentation techniques. In other words, with the maximum potential capacity of both augmentation methods, we would like to investigate which one is more informative to the CNNs classifiers. Therefore, each cell image was rotated by 360°, with an angle step of 18°. Thus, for each input image, 20 rotated-version output images were acquired. However, for the minor (Golgi) class, which has about 1/3 of the average images number of the other individual classes, the rotation angle step was set to 6° to compensate for the difference in image numbers. Then, the horizontal and vertical filliping of each image was added. Thus, the original size of the data is enlarged by a factor of 60 (180 for Golgi). An augmented version of the validation set was created to evaluate the implemented GANs models. For the rest of this paper, we referred to the original training set as tr_orig and the augmented training set as tr_aug. Similarly, val_orig and val_aug are referred to the original and the augmented versions of the validation set, respectively. Table 1 reports the number of images in each class before and after data augmentation.

[1]Download link for the I3A dataset: https://hep2.unisa.it/dbtools.html (Accessed on June 14, 2021)

C. GANs FOR HEp-2 CELL IMAGES SYNTHESIS

Four variants of GANs were trained in an unsupervised manner to learn the visual representation of the HEP-2 cell images. Since HEp-2 data exhibits high intra-classes variations, GAN models were trained individually for each class of the HEp-2 I3A dataset for better representation learning. All architectures of GANs models were set to generate single-channel 64 $\times$ 64 pixels images. The noise vectors z was sampled from random normal distributions ($z \sim \mathcal{N}(0, 1)$). For efficient training, all GANs networks were trained with the augmented training set (tr_aug). The implementation details of each GAN network were summarized in Table 2.

Across all implemented GANs variants, DCGAN was the hardest to train and showed high instability during training even with carefully tuning its hyperparameters. However, Wasserstein-based GANs' losses demonstrated more stable training for this data. The best performing models across all implemented GANs were selected based on the lowest FID scores.

D. GANs EVALUATION METRICS

Since both FID and 1-NN classifier metrics are sample-based methods, a real set S_r was sampled randomly from the augmented version of the validation set (val_aug), whereas GANs' generators were used to synthesize the generated sample sets S_g. The number of samples of each set was fixed to be $5K$ in all evaluation experiments. Each reported metric's value in this paper is an average of five repetitions with random seeds and random sampling from val_aug. The 1-NN classifier metric was implemented using the relatively small HEpNet [19] network (to the final FC layer before the final classifier layer) pre-trained on the augmented training set (tr_aug) to extract data features of length 256.

E. CNNs FOR HEp-2 CELL IMAGE CLASSIFICATION

The optimal GAN models are used to generate two balanced datasets with different sizes composed of 300K and 600K images which are annotated as GAN_300K and GAN_600K, respectively. To evaluate the different augmentation methods, four CNNs that achieved the best performance in the literature of HEp-2 cell image classification were implemented, which are DRINet [18], DCRNet [17], DSRNet [20], HEpNet [19], in addition to the InceptionV3 model [21]. The classification performance of all CNNs were evaluated with different training data variants, particularly, the original-size training data (tr_orig), the classic augmentation data (tr_aug), the Optimal-GANs-generated data (GAN_300K and GAN_600K), and a combination of classic augmentation

TABLE 1. Details of the classic augmentation for each I3A dataset class. tr_orig, val_orig are the training and the validation set before classic augmentation, respectively. tr_aug, val_aug are the training and the validation set after classic augmentation, respectively.

Class	I3A data size	tr_orig (64%)	val_orig (16%)	test set (20%)	Aug. factor	tr_aug	val_aug
Ce	2741	1754	438	549	20×3	105240	26280
Gl	724	463	115	146	60×3	83340	20700
Ho	2494	1596	399	499	20×3	95760	23940
Nu	2598	1662	415	521	20×3	99720	24900
NuM	2208	1413	353	442	20×3	84780	21180
Sp	2831	1811	452	568	20×3	108660	27120
Total	13596	8699	2172	2725	-	577500	144120

TABLE 2. Experimental details of the implemented GANs. MB: mini-batch size, lr: learning rate, and iter: iterations.

Model	MB	Optimizer	Initial lr	β_1, β_2	No. of G iter ($\times 10^3$)	Noise vector size (z)	No. of D iter per one G iter	Other parameters
DCGAN	16	Adam	2×10^{-4}	β_1=0.5, β_2=0.999	100	256	1	-
WGAN	128	RMSProp	5×10^{-5}	β=0.99	100	100	5	clipping= 0.01
WGANGP	16	Adam	1×10^{-4}	β_1=0, β_2=0.9	50	128	10	λ=10
Info-WGANGP	16	Adam	1×10^{-4}	β_1=0, β_2=0.9	100	100	5	vector c=5, λ_1=10, λ_2= 1

and GANs-generated data (tr_aug + GAN_300K and tr_aug + GAN_600K). The pre-defined val_orig set was used to evaluate the training processes, whereas the test set was used for the final CNNs' evaluation. Each model was trained from scratch using the hyper-parameters proposed in the original works, as shown in Table 3.

Two performance metrics were used to evaluate the CNN classifiers; Average Classification Accuracy (ACA), and Mean Class Accuracy (MCA). The ACA is the overall accuracy defined as:

$$ACA = \frac{1}{N} \sum_{i=1}^{N} I(\hat{y}_i = y_i), \tag{7}$$

where N is the number of testing samples, $\hat{y}_i$ is the classifier prediction of the true label y_i, and I is the indicator function. In addition, MCA calculates the mean accuracy of each class, and is defined as:

$$MCA = \frac{1}{M} \sum_{j=1}^{M} CCR_j, \tag{8}$$

where M is the number of classes and CCR_j is the classification accuracy of the j^{th} class. All experiments in this study were implemented utilizing Pytorch framework [47] on a single GPU device (NVIDIA Geforce RTX 2060 SUPER, 8 GB RAM).

V. RESULTS & DISCUSSION

Tables 4 and 5 summarize the FID scores and 1-NN classifier metrics of the implemented GANs across HEp-2 cell classes, respectively. To get an intuition about the lower bounds (annotated in tables as "real"), metrics were computed between two disjoint sets sampled from the real data distribution (val_aug). FID scores reported in Table 4 show that Info-WGANGP outperforms the other GAN models, except for Golgi and Centromere classes in which

TABLE 3. Experimental details of the implemented CNNs. MB: mini-batch size, lr: learning rate.

Model	MB	Optimizer	Initial lr	β_1, β_2	Epochs	Others
DRINet [18]	64	SGD	0.001	$\beta = 0.9$	33	-
DCRNet [17]	128	SGD	0.01	$\beta = 0.9$	80	-
DSRNet [20]	64	SGD	0.01	$\beta = 0.9$	90	weight decay = 0.01
HEpNet [19]	64	Adam	0.0001	$\beta_1 = 0.9$, $\beta_2 = 0.999$	40	-
Incep-V3 [21]	32	SGD	0.01	$\beta = 0.9$	50	-

WGAN and WGANGP achieved slightly better FID scores, respectively. FID scores indicate that introducing mutual information maximization into the WGANGP formulation was beneficial to improve the GAN's capability of learning the underlying modes of the real data distribution. As Info-WGANGP mutually maximized the association between the disentangled visual features of the HEp-2 cell images and the categorical latent vector c, this offered a controllability over the semantic features of the generated images, and as a final reward, enhanced the diversity of the generated images. In the implementation of Info-WGANGP, we uniformly generated HEp-2 cell images from all values of the categorical latent variable c (0, 1, 2, 3 *and* 4), which encourage capturing visual representations of the different successfully learned modes.

In line with the FID metric, Table 5 shows that the Info-WGANGP model achieved the best 1-NN classifier test results (the closest to the chance-level accuracy 50%), indicates the positive impact of integrating the mutual information maximization with WGANGP configuration in improving the capability of GANs to approximate the real data distribution. Another observation is that WGAN achieved the second best 1-NN classifier test scores followed

TABLE 4. FID scores of each GAN model across all data classes (calculated between generated images and real val_aug images). 'real' refers to the FID scores between two disjoint sets of real val_aug images.

Class	real	DCGAN	WGAN	WGANGP	Info-WGANGP
Ce	1.51	23.42	12.80	**7.20**	7.77
Gl	2.69	24.48	**18.37**	40.32	23.39
Ho	2.47	43.05	13.48	12.55	**7.83**
Nu	2.44	29.05	22.65	9.43	**9.31**
NuM	2.83	27.54	13.53	13.18	**8.36**
Sp	2.24	33.99	10.67	10.56	**7.20**

by WGANGP and DCGAN, respectively. In general, Wasserstein-based GANs yielded better results for both metrics compared to the DCGAN. However, a deeper analysis of the other 1-NN classifier measures could reveal interesting observations:

- The real-class accuracies of all experimented GANs are found to be relatively high, indicating that some modes of the real data distribution are not covered by the generator distribution, and hence, all GANs suffers from modes dropping to some extent.
- The generated-class accuracies across all experimented GANs are found to be relatively lower than the other 1-NN classifier metrics. This seems to suggest that within the captured modes, GANs could generate relatively diverse data. However, since the generated-class accuracies are still higher than the chance-level value (50%), there is a presence of some sort of mode collapse too.

These findings demonstrated that the main problem with all adopted GANs is that they tend to drop some modes of the data distribution, in addition to the presence of some sort of collapsing behavior too. This could be reasonable with real-world data exhibiting high visual variances, which is the case of HEp-2 cell images (see Fig. 3). Considering the two metrics, It is clear that even with the best performing GAN, there is still a considerable gap between Info-WGANGP scores and the lower-bound indicating that GANs could not perfectly capture the underlying distribution of the real HEp-2 cell images data, and hence, GANs-generated images are of less diversity compared to the real images.

Fig. 3 depicts some visual examples of GANs-generated images compared to real ones. Real data (upper left group) shows an example of the visual heterogeneity and variances within each HEp-2 data class which give an intuition about the complexity of representation learning task of this data. Even though, GANs-generated images exhibit a reasonable visual similarity with the real data, particularly of Info-WGANGP. As shown in Fig. 3, it is difficult to evaluate the performance of each GAN model merely by visual inspection and thus using quantitative evaluation metrics is necessary for systematic comparison.

For experiments of Phase II, the trained Info-WGANGP generators were used to generate the two balanced datasets, GAN_300K and GAN_600K. The results obtained from all CNNs trained with all variants of the training dataset are reported in Table 6. Results show that training CNNs with standalone GANs-generated data (GAN_300K or GAN_600K) achieved lower classification performance than training them with the classic augmentation data (tr_aug). This may refer to the insufficient diversity of the GANs-generated data, mainly due to mode dropping, even if they could densely generate samples without signs of severe collapsing.

On the other hand, training the CNNs with a combination of the classic augmentation data and the smaller GANs-generated dataset (GAN_300K + tr_aug) achieved a noticeable improvement in the classification performance across all the CNNs under study. Although the training set size was doubled by combining the larger GANs-generated and the classic augmentation datasets (GAN_600K + tr_aug), the obtained results did not show a clear trend of improvement for both ACA and MCA metrics over the classic augmentation dataset (tr_aug). These findings suggest that GANs have a limited capacity for generating informative diverse data. Thus, GANs-generated data could be informative for augmentation up to some size limits while densely sampling from GANs-generators' distributions seem to yield diversity-saturated data, which is found to be nonbeneficial or even have a negative impact on the CNNs performance in some models such as DSR-Net [20] and HEpNet [19]. Therefore, unlike the case of augmenting with real data, further increasing the training data with more GANs-generated images not necessarily yields a corresponding improvement in the classification performance.

It is clear that combining classic augmentation with a limited-size GANs-generated data effectively improved the classification performance of HEp-2 cell images as demonstrated by the results of the combined dataset (GAN_300K + tr_aug) across the five different architectural CNNs. It is important to mention that Table 6 reports the results of our implementations of the CNNs models using hyperparameters' values as proposed in the original works without further tuning. However, for a general overview, Table 7 provides a comparison of the results reported in the original works and our best-achieved results for each implemented CNNs, considering that all works applied almost similar data splitting ratios. Obviously, including GANs-generated data in the training enhanced the discriminability and generalization capability of all the CNNs as shown by the robust accuracy metric MCA results, which indicates the informative impact of the added GANs-generated data. Noticeably, training DCRNet [17] with (GAN_300K + tr_aug) dataset achieved a competitive classification performance for both ACA (98.71%) and MCA (98.89%). Fig. 4 shows the accuracy curves (training and validation) of the DCRNet model across all training data variants during the training process.

However, for further comparison between the real and GANs-generated data, an additional set of experiments were performed by training the DCRNet [17] model five times on

TABLE 5. 1-NN classifier scores of each GAN model across all data classes (calculated between generated data and real data val_aug). 'real' refers to the lower-bound scores computed between two sets of the real data val_aug. 1-NN acc: the overall accuracy, real_class acc: the true positive rate indicating the accuracy among the real data class, gen_class acc: the true negative rate indicating the accuracy among the generated class. The best value is the closest to the chance-level accuracy (50%).

Class	1NN C2ST Metrics	real	DCGAN	WGAN	WGANGP	Info-WGANGP
Ce	1NN acc	0.5014	0.679	0.6299	0.6409	**0.6051**
	real_class acc	0.5326	0.6894	0.6766	0.6687	**0.6231**
	gen_class acc	0.4701	0.6687	**0.5831**	0.613	0.5872
Gl	1NN acc	0.4964	0.7652	0.7619	0.8056	**0.7014**
	real_class acc	0.5272	0.819	0.8482	0.8631	**0.7558**
	gen_class acc	0.4657	0.7114	0.6757	0.7481	**0.6469**
Ho	1NN acc	0.498	0.856	0.7058	0.7628	**0.6711**
	real_class acc	0.5253	0.8656	0.7592	0.7984	**0.7026**
	gen_class acc	0.4706	0.8465	0.6525	0.7272	**0.6396**
Nu	1NN acc	0.4926	0.8587	0.6896	0.6846	**0.6524**
	real_class acc	0.5192	0.8462	0.7466	0.7425	**0.6949**
	gen_class acc	0.466	0.8711	0.6326	0.6267	**0.6099**
NuM	1NN acc	0.4978	0.7644	0.6944	0.7668	**0.6611**
	real_class acc	0.532	0.763	0.7566	0.7974	**0.7031**
	gen_class acc	0.4636	0.7658	0.6322	0.7362	**0.6192**
Sp	1NN acc	0.5073	0.8598	0.71	0.7449	**0.6657**
	real_class acc	0.5314	0.8701	0.7746	0.7928	**0.7125**
	gen_class acc	0.4833	0.8496	0.6453	0.6971	**0.6189**

TABLE 6. Comparison of the classification performances of all used CNNs models across all training data variants. GAN-generated datasets is annotated as GAN_300K and GAN_600K. The approximate size of each training set is written between parentheses. Reported values are in %.

Model	Average Classification Accuracy (ACA)						Mean Class Accuracy (MCA)					
	tr_org (8.7K)	tr_aug (577K)	GAN_300K (300K)	GAN_600K (600K)	tr_aug + GAN_300K (877K)	tr_aug + GAN_600K (1177K)	tr_org (8.7K)	tr_aug (577K)	GAN_300K (300K)	GAN_600K (600K)	tr_aug + GAN_300K (877K)	tr_aug + GAN_600K (1177K)
DCRNet [17]	93.65	98.56	91.08	96.07	**98.71**	98.60	94.30	98.68	92.10	96.46	**98.89**	98.63
DRINet [18]	95.08	98.42	92.80	92.47	**98.64**	98.53	95.62	98.39	93.32	92.69	**98.74**	98.51
HEpNet [19]	90.45	98.42	90.78	89.72	**98.71**	98.38	91.75	98.46	91.97	90.93	**98.81**	98.55
DSRNet [20]	96.14	98.34	84.18	75.30	**98.49**	98.31	96.45	98.50	83.70	74.95	**98.54**	98.38
Incep-V3 [21]	93.00	98.20	94.82	94.53	**98.42**	98.31	93.75	98.36	95.20	94.81	**98.57**	98.37

TABLE 7. Comparison between our best results (achieved by training on tr_aug + GAN_300K dataset) and the results reported in the original works. Reported values are in %.

Model	ACA		MCA	
	Original results	Optimal results	Original results	Optimal results
DCRNet [17]	**98.82**	98.71	98.62	**98.89**
DRINet [18]	98.49	**98.64**	98.37	**98.74**
HEpNet [19]	**98.96**	98.71	98.50	**98.81**
DSRNet [20]	98.42	**98.49**	-	**98.54**

TABLE 8. Classification results of DCRNet model trained on a fixed-size training set (600K) with changing the ratio between real images (collected from tr_aug) and GAN-generated images (collected from GAN_600K) in five steps. Additional random horizontal and vertical shifting were performed on tr_aug images to reach 600K for the first experiment (i.e., 100/0).

Metrics	(real / GAN-generated) ratios in (%)				
	100/0	**75/25**	**50/50**	**25/75**	**0/100**
ACA	98.56	98.49	98.34	98.09	96.07
MCA	98.68	98.62	98.49	98.20	96.46

600K fixed size training set, with changing the (real/GANs-generated) ratio by 25% step at each time, starting from (100% real/ 0% GANs-generated) and ending with (0% real/ 100% GANs-generated). Results are reported in Table 8, which show that as the ratio of the real data decreased, the data diversity in the training set is also decreased and hence the performance of the classifier is consequently deteriorated.

In general, although Info-GANGP demonstrated a superiority in learning the visual representation of the HEp-2 cell images among the other GAN models, the standalone GAN-based augmentation is not as effective as the classic augmentation method as shown in the results of this study. This implying that GANs' capacity of generating diverse data is still limited due to mode dropping and collapsing. Similar findings were observed by [45] who used DCGAN for generating HEp-2 cell images. However, unlike their combined augmentation methods' results, within some data-size limits, using InfoWGANGP generated data as a complementary augmentation method is found to be beneficial to improve the classification performance. These results are consistent with those found in other medical image classification studies used GAN-augmentation methods [31], [32]. As a final remark, the implemented intensive classic augmentation method on the I3A dataset has considerably enlarged the data size to a limit that may eliminate the impact of the information acquired by adding the GANs-generated data in improving the classification performance. Therefore, further investigations on smaller HEp-2 datasets such as

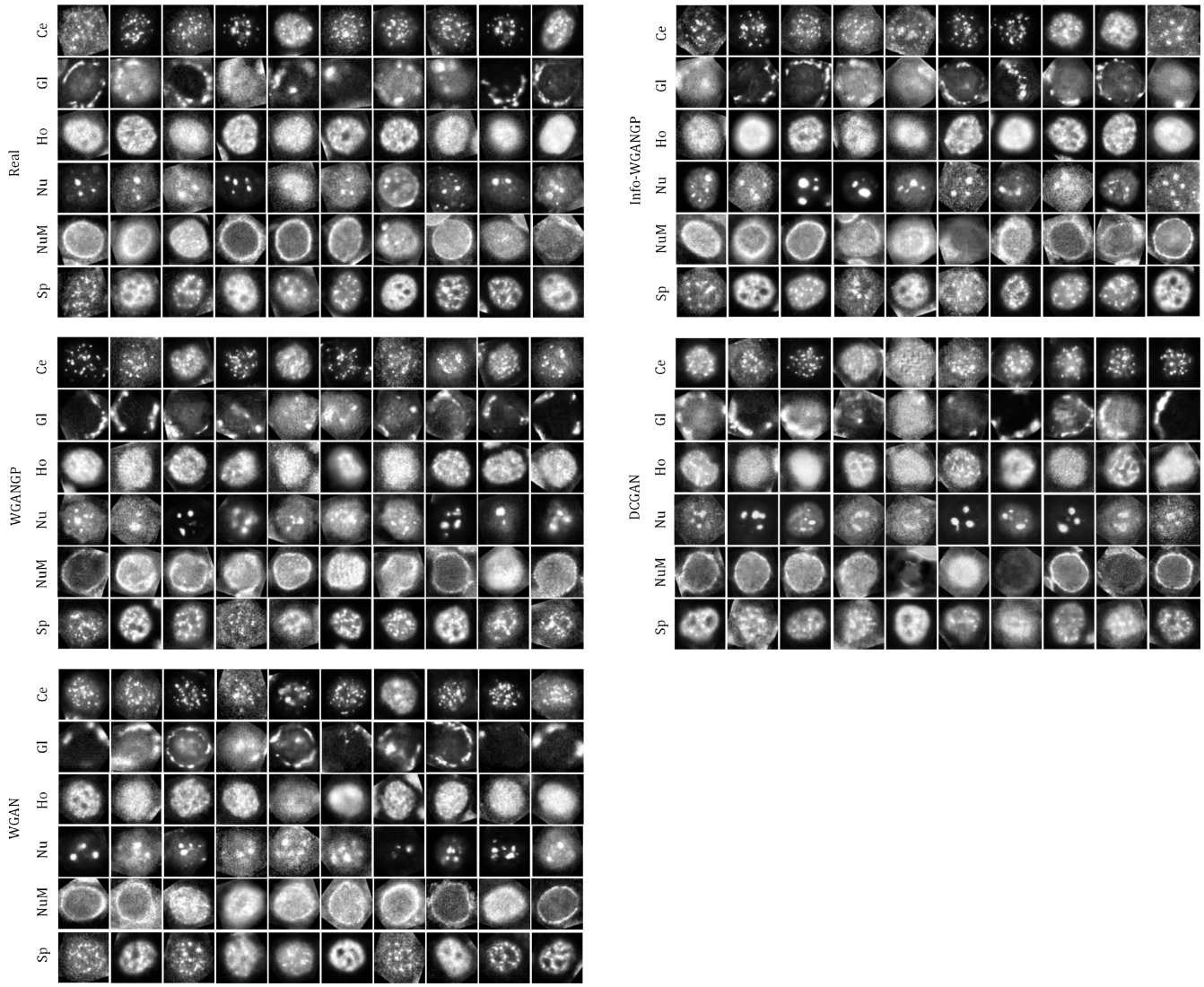


FIGURE 3. Visual comparison between real augmented HEp-2 samples and images generated by the implemented GANs. The image sources are annotated beside each group. Ten randomly selected images are plotted horizontally for each HEp-2 image class. Real data examples of HEp-2 cell images (upper left) show the visual heterogeneity and high variance across classes. GANs-generated images show high similarity to the real data, especially of the Info-WGANGP.

MIVIA[2] and SNPHEp2[3] could help reveal the impact of using GANs augmentation for boosting the classification performance on potentially small training data. Since the sizes of those datasets are approximately 1/10 of that of I3A, GANs pre-trained on the larger I3A dataset could be transferred to the smaller MIVIA or SNPHEp2 datasets. Such investigation is suggested as future work. Moreover, further works are suggested to propose systematic approaches for estimating the optimal size of GANs-generated data to be used as an effective augmentation method.

[2]Download link for the MIVIA (ICPR2012) dataset: https://mivia.unisa.it/contest-hep-2 (Accessed on June 14, 2021)

[3]Download link for the SNPHEp2 dataset: https://staff.itee.uq.edu.au/lovell/snphep2 (Accessed on June 26, 2021)

VI. CONCLUSION

This study provides a detailed investigation of the capabilities of different types of well-known GANs to learn the visual representations of HEp-2 cell images for augmentation purposes. The empirical evaluation metrics used to quantitatively assess the performances of the implemented GANs showed the superiority of WGANGP with a mutual information maximization objective function over the other GAN models under study. Visually, GANs-generated data showed high similarity with the real HEp-2 cell images with no signs of collapsing. However, the evaluation metrics demonstrated that even the best performing GAN suffers from some degrees of mode dropping and collapsing, limiting its capabilities of producing sufficiently diverse data. Adding a limited-size GANs-generated data to the classic augmentation showed a clear improvement in the classification

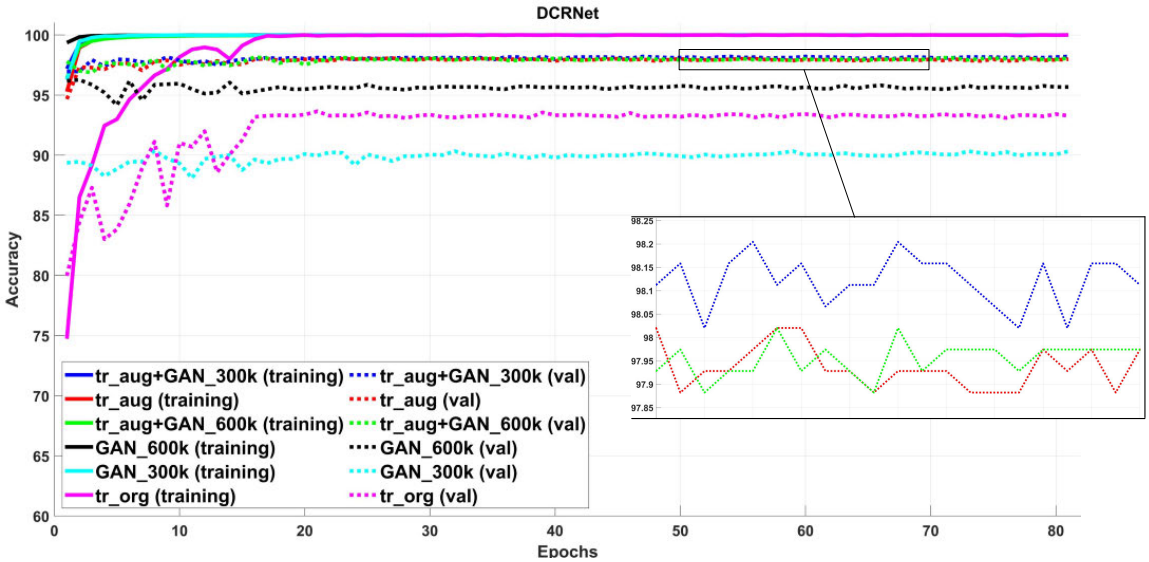


FIGURE 4. Training and validation accuracy curves of the DCRNet model across all training data variants during the training process. A small portion of the overlapped validation curves was zoomed in for better visualization. The figure is better visualized in digital format.

performance across different variants of CNNs architectures achieving a competitive classification accuracy, especially for the DCRNet [17] model. These findings demonstrated the applicability of GANs-generated data for enhancing the generalization performance of CNNs for the HEp-2 cell image classification task.

ACKNOWLEDGMENT

Asaad Anaam would like to express his sincere gratitude to Dr. Mohammed Al-masni, from the Department of Electrical and Electronic Engineering, Yonsei University, Korea, for his valuable advice that contributed to the success of this work.

REFERENCES

[1] P. Hobson, B. C. Lovell, G. Percannella, A. Saggese, M. Vento, and A. Wiliem, "Computer aided diagnosis for anti-nuclear antibodies HEp-2 images: Progress and challenges," *Pattern Recognit. Lett.*, vol. 82, pp. 3–11, Oct. 2016.

[2] P. L. Meroni and P. H. Schur, "ANA screening: An old test with new recommendations," *Ann. Rheumatic Diseases*, vol. 69, no. 8, pp. 1420–1422, Aug. 2010.

[3] P. Foggia, G. Percannella, P. Soda, and M. Vento, "Benchmarking HEp-2 cells classification methods," *IEEE Trans. Med. Imag.*, vol. 32, no. 10, pp. 1878–1889, Oct. 2013.

[4] P. Hobson, G. Percannella, M. Vento, and A. Wiliem, "Competition on cells classification by fluorescent image analysis," in *Proc. 20th IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2013, pp. 2–9. [Online]. Available: https://mivia-web.diem.unisa.it/contest-icip-2013/

[5] P. Hobson, B. C. Lovell, G. Percannella, M. Vento, and A. Wiliem, "Benchmarking human epithelial type 2 interphase cells classification methods on a very large dataset," *Artif. Intell. Med.*, vol. 65, no. 3, pp. 239–250, Nov. 2015.

[6] B. C. Lovell, G. Percannella, A. Saggese, M. Vento, and A. Wiliem, "International contest on pattern recognition techniques for indirect immunofluorescence images analysis," in *Proc. 23rd Int. Conf. Pattern Recognit. (ICPR)*, Dec. 2016, pp. 74–76.

[7] S. Rahman, L. Wang, C. Sun, and L. Zhou, "Deep learning based HEp-2 image classification: A comprehensive review," *Med. Image Anal.*, vol. 65, Oct. 2020, Art. no. 101764.

[8] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 27, 2014, pp. 2672–2680.

[9] X. Yi, E. Walia, and P. Babyn, "Generative adversarial network in medical imaging: A review," *Med. Image Anal.*, vol. 58, Dec. 2019, Art. no. 101552.

[10] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," in *Proc. 4th Int. Conf. Learn. Represent. (ICLR) Conf. Track*, Nov. 2015, pp. 1–16.

[11] A. Martin, C. Soumith, and L. Bottou, "Wasserstein generative adversarial networks," in *Proc. 34th Int. Conf. Mach. Learn.* (Proceedings of Machine Learning Research), vol. 70, D. Precup and Y. W. Teh, Eds. USA: PMLR, Aug. 2017, pp. 214–223. [Online]. Available: https://proceedings.mlr.press/v70/arjovsky17a.html

[12] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville, "Improved training of Wasserstein GANs," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30. I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds. Red Hook, NY, USA: Curran Associates, 2017, pp. 5768–5778.

[13] B. Hu, Y. Tang, E. I.-C. Chang, Y. Fan, M. Lai, and Y. Xu, "Unsupervised learning for cell-level visual representation in histopathology images with generative adversarial networks," *IEEE J. Biomed. Health Informat.*, vol. 23, no. 3, pp. 1316–1328, May 2019.

[14] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel, "InfoGAN: Interpretable representation learning by information maximizing generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, Jun. 2016, pp. 2180–2188.

[15] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "GANs trained by a two time-scale update rule converge to a local nash equilibrium," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, Jun. 2017, pp. 6627–6638.

[16] Q. Xu, G. Huang, Y. Yuan, C. Guo, Y. Sun, F. Wu, and K. Weinberger, "An empirical study on evaluation metrics of generative adversarial networks," 2018, *arXiv:1806.07755*. [Online]. Available: http://arxiv.org/abs/1806.07755

[17] L. Shen, X. Jia, and Y. Li, "Deep cross residual network for HEp-2 cell staining pattern classification," *Pattern Recognit.*, vol. 82, pp. 68–78, Oct. 2018.

[18] Y. Li and L. Shen, "A deep residual inception network for HEp-2 cell classification," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, vol. 10553. Cham, Switzerland: Springer, 2017, pp. 12–20. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-319-67558-9_2

[19] Y. Li and L. Shen, "HEp-Net: A smaller and better deep-learning network for HEp-2 cell classification," *Comput. Methods Biomech. Biomed. Eng.: Imag. Vis.*, vol. 7, no. 3, pp. 266–272, May 2019.

[20] H. Lei, T. Han, F. Zhou, Z. Yu, J. Qin, A. Elazab, and B. Lei, "A deeply supervised residual network for HEp-2 cell classification via cross-modal transfer learning," *Pattern Recognit.*, vol. 79, pp. 290–302, Jul. 2018.

[21] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2818–2826.

[22] Z. Gao, J. Zhang, L. Zhou, and L. Wang, "HEp-2 cell image classification with convolutional neural networks," in *Proc. 1st Workshop Pattern Recognit. Techn. Indirect Immunofluorescence Images*, Aug. 2014, pp. 24–28.

[23] N. Bayramoglu, J. Kannala, and J. Heikkila, "Human epithelial type 2 cell classification with convolutional neural networks," in *Proc. IEEE 15th Int. Conf. Bioinf. Bioeng. (BIBE)*, Nov. 2015, pp. 1–6.

[24] X. Jia, L. Shen, X. Zhou, and S. Yu, "Deep convolutional neural network based HEp-2 cell classification," in *Proc. 23rd Int. Conf. Pattern Recognit. (ICPR)*, Dec. 2016, pp. 77–80.

[25] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. 3rd Int. Conf. Learn. Represent. (ICLR) Conf. Track*, Sep. 2014, pp. 1–14.

[26] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[27] C. Vununu, S.-H. Lee, O.-J. Kwon, and K.-R. Kwon, "A dynamic learning method for the classification of the HEp-2 cell images," *Electronics*, vol. 8, no. 8, p. 850, Jul. 2019.

[28] M. Lu, L. Gao, X. Guo, Q. Liu, and J. Yin, "HEp-2 cell image classification method based on very deep convolutional networks with small datasets," in *Proc. 9th Int. Conf. Digit. Image Process. (ICDIP)*, Jul. 2017, Art. no. 1042040.

[29] C. Vununu, S.-H. Lee, and K.-R. Kwon, "A deep feature extraction method for HEp-2 cell image classification," *Electronics*, vol. 8, no. 1, p. 20, Dec. 2018.

[30] D. Cascio, V. Taormina, and G. Raso, "Deep CNN for IIF images classification in autoimmune diagnostics," *Appl. Sci.*, vol. 9, no. 8, p. 1618, Apr. 2019.

[31] H. Salehinejad, S. Valaee, T. Dowdell, E. Colak, and J. Barfett, "Generalization of deep neural networks for chest pathology classification in X-rays using generative adversarial networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2018, pp. 990–994.

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

[33] C. Baur, S. Albarqouni, and N. Navab, "MelanoGANs: High resolution skin lesion synthesis with GANs," 2018, *arXiv:1804.04338*. [Online]. Available: http://arxiv.org/abs/1804.04338

[34] S. G. Finlayson, H. Lee, I. S. Kohane, and L. Oakden-Rayner, "Towards generative adversarial networks as a new paradigm for radiology education," 2018, *arXiv:1812.01547*. [Online]. Available: http://arxiv.org/abs/1812.01547

[35] B. Lecouat, K. Chang, C.-S. Foo, B. Unnikrishnan, J. M. Brown, H. Zenati, A. Beers, V. Chandrasekhar, J. Kalpathy-Cramer, and P. Krishnaswamy, "Semi-supervised deep learning for abnormality classification in retinal images," 2018, *arXiv:1812.07832*. [Online]. Available: http://arxiv.org/abs/1812.07832

[36] A. Madani, M. Moradi, A. Karargyris, and T. Syeda-Mahmood, "Semi-supervised learning with generative adversarial networks for chest X-ray classification with ability of data domain adaptation," in *Proc. IEEE 15th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2018, pp. 1038–1042.

[37] D. Wang, Z. Lu, Y. Xu, Z. Wang, A. Santella, and Z. Bao, "Cellular structure image classification with small targeted training samples," *IEEE Access*, vol. 7, pp. 148967–148974, 2019.

[38] S. Kazeminia, C. Baur, A. Kuijper, B. van Ginneken, N. Navab, S. Albarqouni, and A. Mukhopadhyay, "GANs for medical image analysis," *Artif. Intell. Med.*, vol. 109, Sep. 2020, Art. no. 101938.

[39] Y. Li and L. Shen, "CC-GAN: A robust transfer-learning framework for HEp-2 specimen image segmentation," *IEEE Access*, vol. 6, pp. 14048–14058, 2018.

[40] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1125–1134.

[41] A. Odena, C. Olah, and J. Shlens, "Conditional image synthesis with auxiliary classifier GANs," in *Proc. 34th Int. Conf. Mach. Learn.*, vol. 70, 2017, pp. 2642–2651.

[42] D. Kastaniotis, I. Ntinou, D. Tsourounis, G. Economou, and S. Fotopoulos, "Attention-aware generative adversarial networks (ATA-GANs)," in *Proc. IEEE 13th Image, Video, Multidimensional Signal Process. Workshop (IVMSP)*, Jun. 2018, pp. 1–5.

[43] K. Gupta, D. Thapar, A. Bhavsar, and A. K. Sao, "Effectiveness of GAN-based synthetic samples generation of minority patterns in HEp-2 cell images," in *Proc. 42nd Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Jul. 2020, pp. 1376–1379.

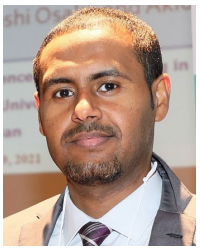
[44] H. Xie, Y. He, H. Lei, J. Y. Kuo, and B. Lei, "Segmentation guided HEp-2 cell classification with adversarial networks," in *Proc. Comput., Commun. IoT Appl. (ComComAp)*, Oct. 2019, pp. 374–379.

[45] T. Majtner, B. Bajić, J. Lindblad, N. Sladoje, V. Blanes-Vidal, and E. S. Nadimi, "On the effectiveness of generative adversarial networks as HEp-2 image augmentation tool," in *Image Analysis*. Cham, Switzerland: Springer, 2019, pp. 439–451.

[46] D. Lopez-Paz and M. Oquab, "Revisiting classifier two-sample tests," in *Proc. Int. Conf. Learn. Represent.*, Toulon, France, Apr. 2017. [Online]. Available: https://iclr.cc/archive/www/doku.php%3Fid=iclr2017:conference_posters.html#wednesday_afternoon and https://dblp.org/rec/conf/iclr/Lopez-PazO17.bib

[47] A. Paszke, S. Gross, F. Massa, and A. Lerer, "Pytorch: An imperative style, high-performance deep learning library," in *Proc. Adv. Neural Inf. Process. Syst.*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds. Red Hook, NY, USA: Curran Associates, 2019, pp. 8024–8035.

ASAAD ANAAM received the B.Sc. and M.Sc. degrees in biomedical engineering and systems from the Faculty of Engineering, Cairo University, Egypt, in 2011 and 2017, respectively. He is currently pursuing the Ph.D. degree with the Graduate School of Interdisciplinary Science and Engineering in Health Systems, Okayama University, Japan. His research interests include biomedical signal and image processing, machine learning, deep learning, and the applications of computer vision in the medical fields.



HANI M. BU-OMER (Graduate Student Member, IEEE) received the B.Sc. and M.Sc. degrees in biomedical engineering and systems from the Faculty of Engineering, Cairo University, Egypt, in 2011 and 2017, respectively. He is currently pursuing the Ph.D. degree with the Graduate School of Interdisciplinary Science and Engineering in Health Systems, Okayama University, Japan. He is the Co-Founder of the Hadhramout Foundation for Invention and Advancement of Science, Yemen. He was the Secretary-General and Treasurer of the Forum of Hadhramout Community, Egypt, from 2015 to 2018. His research interests include programming, system dynamics, medical signal and image processing, machine learning, EEG, and brain–computer interface.

AKIO GOFUKU was born in Japan, in 1957. He received the B.Sc. and M.Sc. degrees in electrical engineering and the Ph.D. degree from Kyoto University, in 1981, 1983, and 1990, respectively. He was an Assistant Professor with the Institute of Atomic Energy, Kyoto University, from 1984 to 1994. He became an Associate Professor at the Faculty of Engineering, Okayama University, in December 1994. He is currently a Professor with the Graduate School of Interdisciplinary Science and Engineering in Health Systems, Okayama University. His research interests include human–machine interfaces, applications of RT and ICT to medical support systems, and spherical motors.