Learning From Limited and Imbalanced Medical Images With Finer Synthetic Images From GANs

XIAOLI QIN\textsuperscript{1}, FRANCIS MINHTHANG BUI\textsuperscript{1,1} (Member, IEEE), HA H. NGUYEN\textsuperscript{1,2} (Senior Member, IEEE), AND ZHU HAN\textsuperscript{2,3} (Fellow, IEEE)

\textsuperscript{1}Department of Electrical and Computer Engineering, University of Saskatchewan, Saskatoon, SK S7N 5A9, Canada
\textsuperscript{2}Department of Electrical and Computer Engineering, University of Houston, Houston, TX 77004, USA
\textsuperscript{3}Department of Computer Science and Engineering, Kyung Hee University, Seoul 446-701, South Korea

Corresponding author: Francis Minhthang Bui (francis.bui@usask.ca)

This research was supported in part by the Chinese Scholarship Council (CSC), in part by the Natural Sciences and Engineering Research Council of Canada (NSERC), and in part by the Natural Science Foundation (NSF) under Grant CNS-2107216 and Grant CNS-2128368.

\textbf{ABSTRACT}  Chest X-ray is a prevalent medical imaging modality for detecting lung diseases. The clinical analysis of X-ray images is usually conducted by radiologists, who represent valuable human resources. In practice, situations with insufficient radiologists to timely analyze large quantities of X-ray data are very common. Accordingly, developing an automated computer-aided lung disease classification system is beneficial to facilitate diagnoses. However, due to restrictions of costs and time, collecting large amounts of accurately labeled X-ray images to train a machine learning based diagnosis system is challenging. Another limitation is the class imbalances present in datasets. Facing these challenges, we investigate the effectiveness of using generative models, particularly generative adversarial networks (GANs), to synthesize new data to tackle the issue of data paucity and class imbalances. To this end, it should be noted that few existing works have studied the effect of generated image quality on the performance of different learning models, particularly in medical imaging. Therefore, the current paper represents one of the first comprehensive investigations into the impact of synthetic image generation on classifier performance, which is empirically elucidated by a comparative analysis of a simpler deep convolutional GAN to a more complex progressive GAN design. Another contribution of this paper is a multi-scale convolutional neural network (CNN) architecture, which can take advantage of image features at different scales for better learning from scratch. Altogether, to verify the robustness of using GANs to augment datasets, we compare various data augmentation approaches, when applied to different network architectures, including transfer learning, learning-from-scratch CNNs, state-of-the-art ResNet, EfficientNet, DenseNet, and the proposed multi-scale CNN. Specifically, testing on two publicly available datasets, the obtained results show that using finer images synthesized from GANs with the proposed multi-scale CNN achieved good classification performance, under a wide range of operating conditions.

\textbf{INDEX TERMS}  Chest X-ray, imbalanced datasets, GANs, CNN, deep learning.

I. INTRODUCTION

The development of computer vision and machine learning algorithms, particularly deep learning methods, has gained increasing attention in medical imaging research. Chest X-ray (CXR) is an indispensable medical imaging modality for screening lung abnormalities. The current research literature related to developing an automated computer-aided diagnosis (CAD) system with CXR images, is unfortunately not well developed and rather lacking. A major underlying reason is the scarcity of large well-labeled CXR datasets to effectively train systems based on supervised machine learning. Indeed, obtaining large labeled medical datasets usually requires costly resources (e.g., professional radiologists are recruited to generate ground truth). Moreover, as a second major challenge, the classes of the available datasets are often

VOLUME 10, 2022  This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/  91663
imbalanced. Using an imbalanced dataset to train a machine learning system directly would lead to biased results. How to make full use of the limited and imbalanced medical datasets to design an effective CAD system remains an open problem.

In this research field, published methodologies can be generally classified into two categories: (1) transfer learning, and (2) training from scratch. A typical transfer learning approach involves using a general pre-trained network as a feature extractor to generate suitable image features, followed by refining the network with a small set of labeled data samples, which are specific to a certain target application domain. Although transfer learning can learn transferable knowledge without a large labeled dataset, the domain shift between the source and target domains, along with overfitting risks, may cause inferior performance on the target domain. By contrast, training from scratch methods are less susceptible to these effects. However, these latter methods often require large labeled datasets for training. To generate a larger dataset, data augmentation is a commonly used strategy. As for the second challenge of class imbalances, resampling the original datasets, either by under-sampling the majority class or over-sampling the minority class, has been the common solution. However, given that available medical datasets are typically limited, over-sampling the minority class is primarily considered to increase the size of datasets.

The most widely used method for augmenting image datasets is geometrical transformation, including flipping, rotation and affine transformations. More recently, generative models, such as generative adversarial networks (GANs), have emerged as potential alternatives for creating new images to augment the training dataset. For instance, Frid-Adar et al. [1] showed that augmenting the dataset with generated medical images from GANs has the potential to improve the performance of a convolutional neural network (CNN), for the liver lesion classification task. The liver lesion datasets in [1] are computed tomography images with a size of $64 \times 64$, which can be identified even by non-experts. However, for CXRs, the image size is often larger and with more details that are difficult to discern without domain-specific radiological knowledge. Therefore, generating CXR images from GANs is considerably more challenging. Apart from methods for directly synthesizing images, the authors in [2] proposed an algorithm called SMOTE to generate synthetic examples in the feature space. They showed that the SMOTE algorithm can improve the performance of traditional classifiers compared to when using under-sampling methods. The authors in [3] applied the SMOTE algorithm with the SVM classifier to reduce the impacts of class imbalances for the task of detecting intestinal contractions. However, handcrafted features and domain-specific knowledge are always required for using SMOTE. By contrast, deep CNNs have the capacity to learn feature representations directly from data, which offer a significant design advantage over traditional machine learning methods.

In our previous paper [4], we explored the effectiveness of using generated images from GANs to augment the limited and imbalanced CXR dataset for automated lung disease classification. Although we showed that the deep CNNs with synthetic images from GANs have a significant potential for medical image applications, some issues and questions remain to be explored. First, how does the image quality (of the synthetic images) affect the overall classifier performance? Second, how should the CNNs be designed, in conjunction with data augmentation, for better feature representation capacity? Third, can all types of learning models (i.e., transfer learning or learning from scratch) perform better as a result of data augmentation using GANs, or is there an inherent behavioral difference between different learning models? It should be noted that, in a more general context, the answers to these questions have thus far been largely elusive in the machine learning literature. And more specifically for the CXR application scenarios, very few works have directly addressed or tackled these questions.

Therefore, as a major extension of our previous work [4], the high-level novelty and contributions of this paper are two-fold: (1) to provide the first comprehensive exposition of the above questions, with observations and insights based on empirical investigations; (2) to propose effective GAN-based solutions that lead to overall system performance improvement, with practical evaluation using CXR datasets. Accordingly, the specific task-based contributions can be summarized as follows:

- We investigate the feasibility of deploying a more complex GAN, namely the progressive growing GAN (PGGAN) [5], to synthesize medical images with higher resolution and finer quality for augmenting imbalanced and limited medical imaging datasets.
- We study the impact of synthetic GAN image quality on the classifier performance, thus quantifying the importance of image quality with respect to data augmentation.
- We propose a multi-scale CNN architecture with the capability of fusing information from image features at different scales. It will be demonstrated that the proposed CNN architecture exhibits good performance, with reduced number of parameters and training time, compared to those from the CNN used in our previous work [4].
- We analyze the performance differences of different types of learning models using augmented images from GANs, in order to provide a better understanding of model selection when applying GANs for addressing the data paucity and class imbalances in medical applications.
- We compare the classification performance of the proposed multi-scale CNN with different state-of-the-art CNNs, using two public CXR image datasets. It will be shown that the imbalance issue can be better remedied by our proposed CNN model, in conjunction with a data augmentation strategy that leverages finer (i.e., higher quality) synthetic images from GANs.
The rest of the paper is organized as follows. Section II discusses related works. Then, the datasets and methods are presented in Section III. Section IV describes the experimental evaluation metrics, along with specific training details of different methods. Next, experimental results and comparisons with various learning methods on real datasets are reported in Section V. Lastly, result discussions and conclusions are presented in Sections VI and VII, respectively.

II. RELATED WORKS

A. CAD METHODS FOR LUNG DISEASE RECOGNITION USING CXR IMAGES

In the context of CAD using CXR images, the authors in [6] proposed an adaptive attention network (AANet), which should be capable of adaptively extracting COVID-19 signs from infected CXR images. Experimental results on three public datasets showed that the proposed AANet outperformed other state-of-the-art methods. However, AANet requires a more complex architecture than a regular CNN. Recently, several works [7], [8], [9] investigated the classification performance of using transfer learning on CXR datasets. Reference [7] showed that fine-tuning a pre-trained Inception V3, trained on ImageNet, can generate promising results on the CXR dataset. However, medical images are often in grayscale and contain a single channel, whereas the inputs of the pre-trained networks trained by natural images often assume 3 channels. Accordingly, in order to fit the pre-trained nets, transforming the grayscale medical image into a three-channel pseudo-color image is always required before feeding the data to the nets. To figure out the difference between fine-tuning a network pre-trained with grayscale version vs. one pre-trained with RGB version, the authors in [8] first trained a ConvNet with the ImageNet in grayscale, and then fine-tuned the ConvNet on two CXR datasets. The results in [8] indicated that using a neural net pre-trained on grayscale ImageNet would gain better accuracy and take less computation time. The diagnosis system in [7] and [8] achieved AUC scores above 0.72 for classifying normal and abnormal CXRs. However, the issue of class imbalances was not addressed. The authors in [9] used four popular CNNs to extract deep features from CXR and computed tomography (CT) images, and subsequently processing these features with different traditional machine learning models to recognize COVID-19 cases. The obtained experimental results showed that linear SVM and multilayer perceptron achieved better disease recognition accuracy than other approaches for both CXR and CT images. It should be noted that this work also did not explicitly address the issue of class imbalances.

B. GENERATIVE ADVERSARIAL NETWORKS

GANs are a sub-category of generative models and were first proposed by Goodfellow et al. [10]. A GAN consists of main two parts: generator and discriminator. The goal of the generator, once properly trained, is to synthesize new fake images that are sufficiently “believable” so as to fool the discriminator. Meanwhile, the discriminator is also trained to have good discernment in differentiating real from fake images. The two modules will be trained until a balance is achieved between the generator and discriminator. As mentioned above, the authors in [1] used a GAN approach to augment the liver lesion dataset, and obtained better sensitivity and specificity. However, it should be noted that the training of GANs is notoriously challenging and time-consuming, often fraught with stability issues. Several improved GANs, such as Wasserstein GAN [11], have been proposed to ensure more stable training and generate images with higher quality. Notably, PGGAN [5] is an improved GAN that can generate high resolution images by using a progressive construction.

C. DATA AUGMENTATION

Conventional data augmentation refers to the process of enlarging the dataset by applying label-preserving transformations. The commonly known data augmentation method for images is geometric transformation. For example, AlexNet [12] employed horizontal reflections and random rotations to augment the data. In the medical imaging domain, similar geometrical transformations were applied for automatically classifying the modality of a medical image [13]. In [14], random elastic deformations of training samples were used to improve the robustness of training to automatically segment neuronal structures in electron microscopic recordings and cells in light microscopic images. Despite the potential of data augmentation methods to augment the dataset effectively, there is still a risk to alter the ground-truth label. Moreover the diversity of generated images from geometric transformation is limited by the genre of the geometrical transform. For example, shifting the images horizontally with different distances makes minor changes. By contrast, GANs have an advantage over the geometric transformations, in generating data with other genre but sharing common feature distributions.

III. DATASETS AND METHODS

A. DATASETS

Two publicly available CXR datasets will be used for demonstrating the effectiveness and general capability of our proposed methods.

1) KAGGLE PNEUMONIA DATASET

The first dataset, henceforth referred to as data1, is available from Kaggle\(^1\). The original dataset was from [7]. The authors in [7] selected CXR images from pediatric patients (from under-one to five-year-old, under routine clinical care from Guangzhou Women and Children’s medical center). The ground truths were labeled by two expert physicians and then checked by a third expert. The original dataset was

\(^1\)https://www.kaggle.com/paultimothymooney/chest-xray-pneumonia
split into training and test sets with sizes of 5,232 and 624, respectively.

In the Kaggle website, the dataset is reorganized into 3 folders (train, validation and test). Each folder contains two categories (pneumonia/normal). In this paper, we use the dataset that was reorganized in the Kaggle website: the training dataset contains 3,875 pneumonia and 1,341 normal images, while the validation set includes 8 pneumonia and 8 normal images. In the test dataset, the numbers of pneumonia and normal images are 390 and 234, respectively. Image size varies from $792 \times 1,560$ pixels to $1,712 \times 1,932$ pixels. Binary classifiers are trained to classify normal and pneumonia CXR images.

2) RSNA PNEUMONIA DETECTION CHALLENGE DATASET

The second dataset, henceforth referred to as data2, is from RSNA. The RSNA Pneumonia Detection Challenge dataset is a subset of the NIH CXR dataset [16] and also available on Kaggle. The RSNA challenge is to detect whether lung opacity exists in a given image by predicting the lung opacity bounding boxes. In essence, this is an objection detection task. Normal lungs are usually filled with air, while lung opacity refers to any lung areas that are replaced with other substances such as fluids and inflammation cells. These symptoms in the lungs manifest in CXRs as those infected areas being opaque.

Although the aim of the challenge is to detect lung opacity by predicting bounding boxes, we only use the available dataset from the website to study classification (normal/lung opacity) as our main objective. There are 26,684 training images in the training datasets, and detailed information about the positive and negative classes is given in the CSV format. Because the competition only provides annotations for the training set, we re-split the training set into train, validation and test sets with ratios of 0.81, 0.09 and 0.1, respectively, for training and evaluation.

There are 3 categories (Lung opacity, Not normal/no lung opacity, Normal) in the RSNA dataset. To ensure the consistency of all the experiments on different datasets, we removed the images that are Not Normal/no lung opacity and simplified the multi-classification task as a binary classification task. Meanwhile there are some images that are mostly in black or white (e.g., see Fig. 1). We screen these non-uniform images out through a black pixel ratio calculation function. After the screening, there are 10,244, 1,141 and 1,243 images in the new training, validation and test datasets respectively. These CXR images are in the DICOM format, and each image size is $1,024 \times 1,024$ pixels.

B. OVERVIEW OF THE PROPOSED CNN ARCHITECTURE

The improved architecture of the proposed CXRs classification CNN is shown in Fig. 2. CNN architectures [1], [17], [18], [19] have been widely introduced for many medical image classification tasks. Considering the small sizes of our medical datasets, and to avoid the problem of overfitting, we previously introduced in [4] a CNN architecture without very complex deep structures. Although the performance of this CNN is competitive, it only takes single-scale inputs. As such, this architecture is likely to lose some useful information of images at other scales.

In fact, multi-scale representation learning has been investigated in the field of computer vision [20], [21], [22]. However, most of these works are related to constructing the scale-invariant feature pyramids from different scale images or image pyramid inputs. As the feature representations obtained through convolution operations may be different at various scales, we use a straightforward method to fuse these features via direct concatenation, like the concatenated blocks shown in Fig. 2.

We will now briefly describe the various operations performed by the proposed CNN architecture. As can be seen from Fig. 2, this architecture exhibits a pattern of re-usable intermediate design blocks, with an emphasis on multi-scale representation learning. First, the architecture takes images of resolution $256 \times 256 \times 1$ as the inputs, then follows with a convolution layer with a kernel size of $3 \times 3$, strides of 2, output channels of 32. Next, the size of images is downscaled to $128 \times 128 \times 1$, and a convolution layer with kernel size of $3 \times 3$, strides of 1, output channels of 1 is applied to the downscaled images. Subsequently, two outputs from the two convolutional layers are fused through concatenation, followed by a max-pooling and dropout layer.

The images are downscaled and convolved once again, to match the shape of the outputs from the dropout layer for concatenation. Then, a new convolutional layer is concatenated with the outputs from the downscaled and convolved images. This is followed by a max-pooling layer. Then, the downsampling and convolution operations are repeated, and concatenated with the outputs of the max-pooling layer. By now, the size of the concatenated outputs is $16 \times 16 \times 65$.

The images are downscaled and convolved once more, to generate outputs with a size of $16 \times 16 \times 1$. Concatenating the two outputs, the latest outputs now have a size of $16 \times 16 \times 66$. As the images with size $8 \times 8 \times 1$ are too coarse to extract meaningful features, we stop the downsampling.

In the end, another 3 convolutional layers, 1 max-pooling layer, 2 fully connected layers with a dropout rate of 0.5 are
The added dropout layers are for reducing the overfitting. Moreover, batch normalization is implemented in the convolutional layers for avoiding the issue of gradient vanishing or exploding. The total number of the parameters of the CNN is approximately 26.3M, which is 0.24M less than that in our previously proposed CNN [4]. In addition, the architecture of the proposed CNN does not have skip connections between different layers. The floating-point operations (FLOPS) and memory access are much less than those in a regular ResNet. This advantage is reflected in the training time comparison of different nets in Table 15.

C. PROGRESSIVE GROWING GANs FOR SYNTHESIZING CXR IMAGES

In our previous work [4], we have investigated the effectiveness of using the Deep Convolutional GAN (DCGAN) to synthesize new CXRs for addressing the issue of training with limited and imbalanced data. Although DCGAN has the potential to generate discernible CXRs, whether the quality of generated images has a significant impact on the final classification performance is still uncertain.

To understand the effects of the quality of generated CXRs on the classifiers, we will explore a more complex GAN architecture, known as progressive growing GAN (PGGAN) [5]. PGGAN demonstrated the capacity to produce natural images of unprecedented quality, such as CELEBA [23], LSUN bedroom [24] and CIFAR10 [25] images. In particular, the authors in [5] created a high-quality version of CELEBA dataset at a resolution of $1,024 \times 1,024$, which had never been achieved in the previous GAN literature. Given that the quality of images generated by PGGAN is generally higher than that achieved by DCGAN, we will also apply PGGAN to synthesize finer (high-quality) CXR images in this work.

Similar to other GANs, PGGAN has two main components: discriminator and generator. However, the specific architectures of these two components are substantially more complex, e.g., compared to the counterparts in DCGAN. Figs. 3 and 4 show the growing architectures of the discriminator and generator with increasing resolutions.

From Fig. 3, we can see that starting with low-scale $16 \times 16$ images, the operation $\text{from-rgb}$ is a $1 \times 1$ convolution to transform a RGB/Gray image to feature maps, while $\text{to-rgb}$ is the inverse of $\text{from-rgb}$. Downscale2d
X. Qin et al.: Learning From Limited and Imbalanced Medical Images With Finer Synthetic Images From GANs

FIGURE 3. Discriminator & Generator architectures at resolution of 16 × 16.

and upscale2d are operations for halving and doubling the image resolution through average pooling and nearest neighbor filtering, respectively. The initial discriminator has an input size of 16 × 16 × 1 and comes with two convolutional layers with a kernel size of 3 × 3 and fully connected layers. The activation function used for each layer is leaky-ReLU. After completing the low-resolution training, layers are progressively added to the networks for higher resolution training.

As illustrated in Fig. 4, new convolutional layers are faded into the network for the training of images at the resolution of 32 × 32. To make sure the transition happens smoothly, weight \( a \) is introduced to control architectures. As weight \( a \) grows linearly from 0 to 1, the resolution of the training images gradually increases from 16 × 16 to 32 × 32. This process is repeated for higher and higher resolutions, until the resolution of the generated images is 256 × 256, as shown in Fig. 5.

In essence, the training of GANs involves minimizing the distance between the real distribution and generated distribution. However, using the loss formulation presented in [10] is not always stable for training. To stabilize the training process, WGAN-GP loss [26] is adopted in the PGGAN architecture. We denote the generator and discriminator as \( G \) and \( D \), respectively. If the input image is represented as \( x \), the output of the discriminator \( D \) would be \( D(x) \). The generator \( G \) gets initialized with a prior input distribution (such as uniform distribution, standard normal distribution), defined as \( p_z(z) \), and outputs the generated image as \( G(z) \), which has the distribution \( p_g \). The real images have the distribution \( p_r \).

Given the above setup, the training of GANs is tantamount to playing a minimax game between \( G \) and \( D \), in order to get the optimum for \( p_g = p_r \). It is challenging for the discriminator and generator to attain the optimum synchronously in practice, and discontinuous distributions of \( p_r \) and \( p_g \) will lead to the training of GANs being massively unstable [27]. Instead of using the Jensen-Shannon (JS) divergence, the Wasserstein-1 distance [26] exhibit more desirable properties for stable training of GANs. Moreover, to meet the condition of weight constraints, the gradient penalty is added to the Wasserstein loss function. Accordingly, for PGGAN training, our loss functions are defined as:

\[
\mathcal{L}^G_G(\tilde{x}) = \min_G(-\mathbb{E}_{\tilde{x} \sim p_{\tilde{x}}}[D(\tilde{x}))],
\]

(1)

for the generator, and,

\[
\mathcal{L}^D_D(x, \hat{x}, \tilde{x}) = \min_D(\mathbb{E}_{x \sim p_x}[D(x)] - \mathbb{E}_{\tilde{x} \sim p_{\tilde{x}}}[D(\tilde{x})]) + \lambda \mathbb{E}_{\hat{x} \sim p_{\hat{x}}}[\|\nabla_\hat{x}D(\hat{x})\|_2^2 - 1]^2],
\]

(2)

for the discriminator, where \( \lambda > 0 \) is a balancing coefficient, and \( \hat{x} \) is the combination of \( \tilde{x} \) and \( x \):

\[
\hat{x} = \alpha x + (1 - \alpha)\tilde{x},
\]

(3)
with $\alpha$ being a random variable drawn from a uniform distribution, i.e., $\alpha \sim \mathcal{U}[0, 1]$, and $\bar{x} = \mathcal{G}(z)$.

To determine suitable parameters for the above equations (2) and (3), the following training and selection strategies are applied. For training the PGGAN architecture, the input of the generator is drawn from the standard normal distribution with a length of 256. All the input images of the discriminator are scaled to the range $[-1, 1]$. The slope of the Leaky-ReLU activation function is 0.2. Weights are initialized with a standard normal distribution and normalized. Adam optimizer [28], with parameters $\beta_1 = 0.0$ and $\beta_2 = 0.99$, is used for updating the weights. The penalty parameter $\lambda$ is set to be 10. In effect, these parameters are consistent with the training configurations described previously in [26].

**IV. EXPERIMENTS**

Given the proposed methods, we conduct a set of experiments on the two datasets described in Section III. To investigate whether the quality of images generated from GANs for augmenting an imbalanced dataset induces significant effects on the classifiers, we will compare three data augmentation strategies: (I) synthetic images from PGGAN, (II) generated images from DCGAN, and (III) images from classical geometrical transformations.

Although the synthetic images from PGGAN are known to exhibit finer subjective visual quality, objective metrics should still be evaluated for a more repeatable quantification. To this end, it is noted that quantitative metrics widely used for measuring the performance of GANs include the Inception Score (IS) [29] and the Frechet Inception Distance (FID) [30]. The IS calculates the statistics of the outputs of the Inception v3 network [31], pre-trained on ImageNet, when fed with the generated image samples. However, the IS score does not use the statistics of real images, nor compare to the statistics of synthetic images. Furthermore, a study [32] shows that applying the IS score to GANs trained on datasets other than ImageNet may result in misleading results. Compared to IS, FID can be viewed as an improved metric, which measures the Wasserstein-2 distance between real and synthetic images. Another relevant objective metric is the multi-scale structural similarity (MS-SSIM) [33], which measures the perceptual similarity of images. The MS-SSIM value range is $[0,1]$. A higher MS-SSIM value implies that the images being compared are more similar perceptually. Accordingly, in this paper, we will use FID and MS-SSIM as the two metrics for evaluating the quality of synthetic images from generative models.

In addition to analyzing the influence of synthetic image quality on the classifiers, we also compare the proposed classifier that is trained from scratch with other state-of-the-art methods, including transfer learning [7], ResNet [15], DenseNet [34] and EfficientNet [35]. We set the classical data augmentation method based on geometrical transformation as the baseline for comparisons.

The specific operations of geometrical transformation include random rotation, flipping and shifting. As for the transfer learning method, the pre-trained Inception V3 on ImageNet is adopted for capturing features of the images. One convolutional layer and two fully connected layers are added to the feature layer that is on top of the pre-trained net for re-training on the CXR datasets. ResNet is composed of residual blocks, which can learn the residual function, and is well known to exhibit excellent performance on classification tasks [15], [36], [37], [38]. Furthermore, DenseNets [34] are reported to achieve even lower error rates than ResNets on several benchmark datasets. In addition, EfficientNets [35] are well known for their state-of-the-art accuracy on ImageNet, with smaller model size and faster inference speed than many existing CNNs.

In our experiments, to avoid the problem of overfitting, we utilize a classical ResNet-18 with 18 layers for training. Likewise, the specific variants of DenseNet-121 and EfficientNet-B0 are used for comparative purposes. All these experiments are executed on a single GeForce RTX 2080 GPU with a memory size of 8GB.

**A. EVALUATION METRICS**

Although accuracy (abbreviated to Acc, see equation (4)) is a metric that has been commonly used to measure the performance of classifiers, it weighs the influences of classes on the number of instances. In a nutshell, classifiers tend to show a bias to the majority class in an imbalanced dataset [39]. Hence, for a more complete analysis, other metrics commonly used for classification with imbalanced datasets will also be considered, including specificity, sensitivity, precision, F1-score, G-mean, AUC score and the area under precision-recall curve AUPRC [39], [40]. These metrics are defined as follows:

\[
\text{Acc} = \frac{TP + TN}{TP + FN + TN + FP}, \quad (4)
\]

\[
\text{Precision} = \frac{TP}{TP + FP}, \quad (5)
\]

\[
\text{Recall} = \frac{TP}{TP + FN}, \quad (6)
\]

\[
F1 = \frac{2 \cdot \text{precision} \times \text{recall}}{\text{precision} + \text{recall}}, \quad (7)
\]

\[
\text{Specificity} = \frac{TN}{TN + FP}, \quad (8)
\]

\[
\text{G-mean} = \sqrt{\text{specificity} \times \text{recall}}, \quad (9)
\]

where TP, TN, FP and FN are true positives, true negatives, false positives and false negatives respectively.

Among those evaluation metrics, the F1-score (7) is the harmonic mean of the precision and recall. Precision (5) is the rate of the correct positive predictions among positive predictions. Recall (6) is also known as sensitivity, which reflects the true positive rate among the class instances. Specificity (8) refers to the true negative rate, which can be interpreted as the ability of the classifier to correctly reject healthy patients without a condition. F1-score indicates the trade-off between precision and recall. The G-mean (9) measures the
balance between the positive accuracy and negative accuracy. F1-score and G-Mean can provide more insights than accuracy when dealing with the classification of imbalanced datasets.

AUC refers to the area under the Receiver Operating Characteristic (ROC) curve [41]. ROC is plotted with the true positive rate (TPR) against the false positive rate (FPR), and reflects the performance of a classifier at different discrimination thresholds. AUC can be viewed as an aggregation measure of performance over various classification thresholds. AUC ranges in value from 0 to 1. The higher the AUC values is, the better the classifier can distinguish the two classes.

Although AUC and ROC curve can be powerful ways for evaluating the classifier’s performance, ROC curve may provide an overly-optimistic view of the classifier’s performance with highly-imbalanced datasets [42]. Instead, the Precision-Recall (PR) curves can be more suitable for imbalanced datasets because the PR curve is created by plotting precision against recall. Many recent works [40], [43], [44] use the area under the PR curve (AUPRC) for comparison.

**B. DATA PREPROCESSING AND ANALYSIS**

In our experiments, the two original datasets have images of varying sizes. The first dataset, data1, is in JPEG format and the second one, data2, is in DICOM. First, we read, normalize, and convert these images into arrays of data type uint8. Then, to make the image size match the input requirements of the networks, we resize these images into $256 \times 256 \times 3$ using the bilinear interpolation. In addition, we center-crop these images for removing regions that are not in the vicinity of the lung. Since the input size of the Inception v3 net is $299 \times 299 \times 3$, we resize the images again to match this input specification. For other networks, we turn the images into grayscale with a size of $256 \times 256 \times 1$ to facilitate the training process.

Imbalance Ratio (IR) is the commonly used metric to quantify the imbalance level of a dataset. The IR is defined as:

$$IR = \frac{N_{\text{majority}} + 1}{N_{\text{minority}} + 1},$$

(10)

where $N_{\text{majority}}$ and $N_{\text{minority}}$ are the sizes of the majority and minority classes, respectively. When IR = 1, the dataset is balanced. When IR > 1, the larger the IR is, the more imbalanced the dataset is. Table 1 lists the class distributions of the two datasets we use for evaluation.

Table 1 shows that these two datasets have different levels of imbalance ratio and sizes. The oversampling method of geometrical augmentation on the minority class would serve as a benchmark for addressing the imbalance issue. Subsequently, oversampling methods of synthesizing instances from DCGAN and PGGAN on the minority class will be comparatively studied, in order to analyze the effectiveness of using synthetic images to balance the class distributions.

| Dataset          | Samples | Minority Samples | Majority Samples | Imbalance Ratio |
|------------------|---------|------------------|------------------|-----------------|
| data1 training   | 5216    | 1311             | 3875             | 2.89            |
| data1 test       | 624     | 234              | 390              | 1.67            |
| data2 training   | 10224   | 3961             | 6283             | 1.59            |
| data2 test       | 1243    | 485              | 758              | 1.56            |

After balancing the datasets, different classifiers will be tested, as discussed previously.

**C. GEOMETRICAL DATA AUGMENTATION**

In these experiments, we only balance the class distributions of the training datasets. The validation and test datasets are unchanged. The geometrical transformation used are random rotation with an angular range of [0, 10], horizontal flipping, horizontal shifting and vertical shifting within a range of $[-5, 5]$. For the first dataset (data1), the minority class is the normal one. We apply the classical augmentation method to generate 2,560 images to balance the classes. For the second dataset (data2), the minority class is the Lung opacity type. We generate 2,080 images to adjust the imbalance ratio to 1.0 through the same geometrical transformations.

**D. TRAINING DETAILS OF DCGAN TO GENERATE IMAGES**

The DCGAN architecture employed is the same as the one reported in our previous work [4]. The training samples of the minority class are fed into the DCGAN to implement unsupervised training. Since the sample sizes of the two datasets are different, the corresponding training parameters also need to be varied accordingly for good performance. For the first dataset: the training epoch is 100, the learning rate is 0.001, and the parameters for the Adam optimizer are $\beta_1 = 0.5$, $\beta_2 = 0.999$. For the second dataset: the training epoch is 300 and the learning rate is 0.0001, and the parameters for the Adam optimizer are the same as those for the first dataset. The batch size is 64 for both datasets. Overall, 2,560 and 2,080 images are generated for the two datasets, respectively, to rebalance the two classes.

**E. TRAINING DETAILS OF PGGAN TO GENERATE IMAGES**

The architectures of PGGAN are shown in Fig. 3 to Fig. 5. The constituent network modules are progressively growing when feeding images from low resolution to high resolution gradually. The lowest resolution of the images used for both datasets is $16 \times 16 \times 1$. The highest resolution for both datasets is $256 \times 256 \times 1$. As for the DCGAN case, only images of the minority class are subjected to training. The training procedure can be decomposed into a sub-training and a resolution transition training stage for different scales. In particular, the sub-training epoch value for the first dataset is 300 at the resolution of $16 \times 16 \times 1$, and the resolution transition training epoch value is 200 for training images with the resolution from $16 \times 16 \times 1$ to $32 \times 32 \times 1$. 

TABLE 1. Summary of the distributions of two datasets.
gradually. Therefore, the total training epoch value for data1 is \((300 + 200) \times 4 + 300 = 2,300\). For the second dataset, the sub-training epoch and resolution transition epoch values are both 100. As such, the total training epoch value for the second dataset is \((100 + 100) \times 4 + 100 = 900\). Moreover, the learning rate selected for the resolutions other than 256 × 256 × 1 is 0.001, whereas for the resolution of 256 × 256 × 1 it is slower at 0.0005. The training batch size is varied from 34 to 17 with the increasing resolution. After training is completed, images generated from PGGAN are added to the minority class training samples, making the imbalance ratio close to 1.

**F. TRAINING DETAILS OF CLASSIFIERS**

For the proposed multi-scale CNN classifier, the weights are initialized with the Xavier method to avoid the issue of weight explosion or diminishing. The batch size is 64, and the learning rate is 0.005. Adam optimizer is adopted to train the network. The training epoch is 20, and early stopping is applied when the validation error does not decrease.

For the transfer learning method, Adam optimizer is also applied for updating the weights. The best learning rate after several trials is 0.0004 ∼ 0.0005. The training epoch values for data1 and data2 are 150 and 250, respectively. Early stopping is also applied.

For the ResNet, the optimizer is the Momentum optimizer with a momentum of 0.95, with the learning rate decaying by a factor of 0.5 every 5 epochs. The exponential moving average is applied to the training loss with a decay factor of 0.8. The initial learning rate is 0.01, and the training epoch value is 20 with a batch size of 16.

For the DenseNet, the optimizer is Adam with a learning rate of 0.001, and the training epoch is 20 with a batch size of 32.

For the EfficientNet, we also use the Adam optimizer with a learning rate of 0.001. The training epoch is 30, and the batch size is 32.

Early stopping is applied to all models.

**V. RESULTS**

**A. QUALITATIVE AND QUANTITATIVE ANALYSES OF GENERATED IMAGES**

Visual quality comparisons of sample generated images from different GANs are shown in Fig. 6. The quantitative analyses of the generated images from different GANs are shown in Table 2.

Observing Fig. 6, we can see that images generated from PGGAN have a finer perceptual quality. Figs. 6 (c) and (f) have clearer spine and lung tissue shapes than Figs. 6 (b) and (e). In addition from Table 2, we can see that, for both datasets, the generated images from PGGAN always have higher quality than those from DCGAN in terms of the quantitative metrics.

**TABLE 2. FID and MS-SSIM comparison of DCGAN and PGGAN: Lower FID and higher MS-SSIM are better.**

| Datasets | GAN-type | FID | MS-SSIM |
|----------|----------|-----|---------|
| data1    | DCGAN    | 293.262 | 0.392   |
|          | PGGAN    | 132.734 | 0.434   |
| data2    | DCGAN    | 211.69  | 0.415   |
|          | PGGAN    | 187.83  | 0.424   |

**B. COMPARISONS OF CLASSIFICATION METHODS USING DIFFERENT DATA AUGMENTATION STRATEGIES**

Extensive experimental results are presented in this section in order to provide insights to the following major issues: (A) whether using images from GANs would be more competitive than when using images from classical data augmentation methods, in order to tackle the imbalanced classification task; (B) whether the higher quality images generated from complex GANs would lead to a leap in performance improvement for all classifiers; and (C) whether all different classifiers can coordinate well with different data balancing strategies.

1) **RESULTS ON Data1 FOR DIFFERENT CLASSIFIERS USING DIFFERENT DATA AUGMENTATION METHODS**

Tables 3–8 summarize the classification results of different classifiers using different data balancing methods on the first dataset (data1). From these tables, we can observe that after balancing the classes, the F1 score, G-mean, AUC, and AUPRC are all improved for any of the classifiers. Notably, our proposed multi-scale CNN with generated images from PGGAN receive the highest F1-score, G-mean and AUC. Therefore, with respect to issue (A) identified above, these results show that augmenting the minority class with generated images from GANs are mostly beneficial for the classification on data1.

Next, with respect to issues (B) and (C), the findings are not as clear-cut, and depend significantly also on the classifier being used. For instance, although images generated from PGGAN have finer visual quality, improvements on classification results are varying for different classifiers. Accordingly, in the following, we will examine the effects of different balancing methods on each classifier individually, in order to draw more precise conclusions.

- **Proposed multi-scale CNN** (Table 3): for this classifier, the difference in results between using generated images from DCGAN and using those from PGGAN is noticeable. For instance, the F1-score of the multi-scale CNN with images from DCGAN is 0.919, whereas it is improved to 0.925 when using PGGAN. Other metrics, such as G-mean, AUC, AUPRC, using PGGAN are all better than those using DCGAN and geometrical transformations.

- **Previously proposed CNN** [4] (Table 4): this classifier exhibits a similar performance profile as for the multi-scale CNN. For example, the G-mean, AUC, AUPRC,
F1-score using generated images from PGGAN are, respectively, 0.005, 0.002, 0.007 and 0.004 higher than the corresponding results from DCGAN.

- **ResNet** (Table 6): the results also show that augmenting with images of higher quality leads to salient improvements. Specifically, the AUPRC using generated images from PGGAN is 0.965, which is 0.009 higher than that from classical augmentation method, and 0.011 higher than that from DCGAN. Likewise, the corresponding AUC is 0.944, which is 0.017 and 0.015 higher than those from DCGAN augmentation and classical augmentation, respectively.

- **DenseNet** (Table 7): the results also show that DenseNet with augmented images from the PGGAN achieves the best performance results. For example, the G-mean value is 0.856, which is considerably higher than those from other augmentation methods. The same performance behavior is also observed for the F1-score, AUC and AUPRC scores, with the advantage attributed to the finer PGGAN augmentation method.

- **EfficientNet** (Table 8): this scenario also presents similar performance results. Augmenting with images from DCGAN and PGGAN is shown to deliver better performance metrics than those from classical augmentation methods. In particular, the EfficientNet with finer quality generated images from PGGAN has much higher F1-score, G-mean, AUC score values than those from other data augmentation methods.

- **Transfer learning** (Table 5): here, the differences between various augmentation methods are not significant. Indeed, the AUC scores of augmentation with images from geometrical transformation, DCGAN and PGGAN are nearly identical, being respectively: 0.928, 0.931, 0.930. Similarly, the corresponding AUPRC scores are 0.944, 0.948 and 0.951. These results are all close to 0.946, which is attained without any augmentation.
The above observations show that successful data balancing actions can deliver significant performance improvement, but this is also contingent on being coupled with a suitably designed classifier. In particular, performance improvement is observed for the newly proposed multi-scale CNN, the previously proposed CNN, as well as for ResNet, DenseNet and EfficientNet. However, it appears that transfer learning is not particularly responsive to the imbalance problem. A possible underlying reason is the inherent domain shift.

It should be noted that the transfer network was pre-trained on ImageNet, which is composed of mostly natural images. And since natural images are very different from CXRs perceptually, the features learned from the pre-trained net are likely shallow and general. These general features are not sufficiently representative for the classifiers to differentiate each class effectively. Accordingly, this domain shift leads to the above rather lackluster performance difference for transfer learning.

2) RESULTS ON Data2 FOR DIFFERENT CLASSIFIERS USING DIFFERENT DATA AUGMENTATION METHODS

To show repeatability, while demonstrating the generalizability of our proposed method, a second series of investigations are also conducted on the second dataset (data2), which has a larger size compared to data1. Tables 9–14 show the performance results on the second dataset (data2), with the same respective classifiers as in the previous section.

It can be observed from the tables of results that all classifiers exhibit higher F1-score, G-mean and AUPRC, when augmented with synthetic images from PGGAN than from other methods. Notably, the proposed multi-scale CNN with generated images from PGGAN still achieve the highest F1-score and G-mean values. In general, except for transfer learning, all classification models achieve better results when augmented using GANs than when using classical geometrical transformations. Furthermore, the performance metrics using a more complex PGGAN are unanimously better than those using a simpler DCGAN. These results are consistent with the conclusions gleaned from testing on data1 previously.

C. TRAINING TIME AND NUMBER OF PARAMETERS AMONG CLASSIFIERS

Tables 3 and 9 show that the proposed multi-scale CNN achieves good performance results, even without any data augmentation method. To investigate whether the proposed architecture has a reasonable training cost, we compare the per-epoch training time and the number of parameters for the various classification models. All models are trained on the same hardware. The resulting values are summarized in Table 15. Notably, the multi-scale CNN has about 26.33M parameters, while only imposing an average per-epoch training runtime of 4.3s on images with a resolution of 256 × 256 × 1. This shows that the multi-scale CNN is more competitive than other models in terms of both training time and performance.

VI. DISCUSSIONS

This work focuses on exploring the effects of synthetic image quality, as generated from GANs, on different classifiers, in order to address the challenge of classification on limited and imbalanced CXR datasets. The performance
differences of different types of learning models, with image datasets augmented using GANs, are also quantified. Built on our previous work [4], a multi-scale CNN classifier is proposed to capture richer image features, thus demonstrating improved performance and training efficiency. These advantages were evaluated and confirmed on two public CXR medical datasets.

For experimental evaluation, two variants of GANs are investigated: a simpler DCGAN and a more complex PGGAN. In particular, PGGAN should allow for generating finer high-quality CXR images. Two objective metrics, FID and MS-SSIM, are used to assess the image quality. The results in Table 2 and Fig. 6 confirm that the PGGAN-generated CXR images indeed exhibit better quantitative metrics and finer perceptual qualities.

For assessing the classification performance on imbalanced CXRs, the baseline for comparison is the case of using classical geometrical transformation methods to augment the minority class in the data. Subsequently, more innovative augmentation methods using DCGAN and PGGAN are designed and implemented. Results from Table 3 to Table 12 indicate that these methods are indeed more effective in improving the classification performance on imbalanced CXR images. Furthermore, generated images with higher quality are generally conducive to better classification results.

These investigations are extensively evaluated on several types of classifiers: a recently proposed CNN, a proposed multi-scale CNN, a ResNet, a DenseNet, an EfficientNet (learned from scratch), and a transfer learning method.

Without any data balancing methods, the CNNs show better performance than ResNet and transfer learning. After augmentation using GANs, the classification performances of the CNNs, ResNet, DenseNet and EfficientNet are all improved, notably with finer generated images from PGGAN. However, the performance improvement of the transfer learning method is not as significant, which is likely due to the domain difference between datasets used by the pre-trained network and the CXR datasets in the target domain.

The proposed multi-scale CNN is strategically designed to fuse features from different image scales. By reducing the number of convolution filters in our previous CNN and exploiting the features at different image scales, the proposed design is highly competitive, demonstrating even better classification performance than other classifiers on different datasets without any augmentation. After augmented with generated images from PGGAN, our proposed classifier architecture achieves the highest F1-score and G-mean on
TABLE 14. Classification results of EfficientNet [35] using different data augmentation methods on data2.

| Evaluation metric | EfficientNet [35] |
|-------------------|------------------|
|                   | No Aug | Geo Aug | DCGAN Aug | PGGAN Aug |
| Acc               | 0.910  | 0.891   | 0.901     | 0.905     |
| Recall            | 0.842  | 0.814   | 0.833     | 0.833     |
| Precision         | 0.897  | 0.904   | 0.906     | 0.916     |
| F1-score          | 0.868  | 0.857   | 0.868     | 0.873     |
| Specificity       | 0.935  | 0.945   | 0.945     | 0.951     |
| G-mean            | 0.888  | 0.877   | 0.887     | 0.890     |
| AUC               | 0.943  | 0.949   | 0.950     | 0.959     |
| AUPRC             | 0.924  | 0.934   | 0.941     | 0.950     |

TABLE 15. Number of parameters and average per-epoch training runtime (RT) for different models.

| Method                  | Input Size | Param, M | Avg. RT, s |
|-------------------------|------------|----------|------------|
| Transfer learning [7]   | 229 x 229 x 3 | 2.92     | 67         |
| ResNet [15]             | 256 x 256 x 1 | 2.75     | 35.2       |
| DenseNet [34]           | 256 x 256 x 3 | 7.04     | 63.0       |
| EfficientNet [35]       | 256 x 256 x 3 | 4.05     | 39.0       |
| CNN [4]                 | 256 x 256 x 1 | 26.58    | 5.7        |
| Multi-scale CNN         | 256 x 256 x 1 | 26.33    | 4.3        |

both considered datasets. From a complexity perspective, it also outperforms other models in terms of the training runtime, as explained in Sec. V-C.

In terms of limitations for the proposed methods, we identify several outstanding issues. First, to generate images with finer quality using PGGAN, it usually requires significantly more computational resources and longer training time compared to coarser methods, e.g., involving DCGAN and geometrical transformations. This is an important trade-off to be balanced, especially when resources may be constrained. Second, although experimental results show that augmenting with finer quality generated images from PGGAN can boost the performance of most classifiers for addressing the issue of class imbalances and data paucity, this augmentation strategy is not a categorical guarantee towards success.

In fact, there are practical limitations to be considered when applying this methodology. For example, if less than a few hundreds of images are available, the training of GANs may fail to converge to a stable state, such that the generated images do not consistently resemble the desired real images. One of the possible underlying reasons is that the available training samples are too few, and thus are insufficient to represent the real positive CXR image distributions. Accordingly, if we only have very limited training samples, we cannot reasonably expect to achieve good-quality generated images from GANs directly.

VII. CONCLUSION

In this paper, we investigate how the quality of GAN synthetic images impacts the classification performance on limited and imbalanced CXR datasets. Empirical results show that generated images with finer quality can noticeably improve the performance of most classifiers, except transfer learning. In terms of contributions, we also propose a highly competitive multi-scale CNN classifier that, when suitably coupled with datasets augmented using PGGAN, achieves excellent results on two public CXR datasets, while incurring the lowest training time compared to other competitors. Our findings indicate that data augmentation using GANs is beneficial for addressing the imbalance problem in medical image analysis, notably outperforming classical geometrical transformation methods. The finer the quality of the generated images is, the better the performance of the classifiers (particularly for models trained from scratch) will be.

With respect to future directions, several promising topics for extending this work are worth highlighting. In the present paper, the evaluation of GAN effectiveness for data augmentation has focused primarily on a binary classification task. To diversify the applicability of this framework, the feasibility of using GAN synthetic images for multi-class disease classification tasks remains to be explored. And in this context, we should also consider the challenge of alleviating domain shift, so that transfer learning methods can perform effectively for various operating conditions. Last but not least, the potential of using explainable artificial intelligence (XAI) algorithms to discover disease-related information via the CNNs should be another worthy objective to be pursued.

REFERENCES

[1] 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,” Neurcomputing, vol. 321, pp. 321–331, Dec. 2018.
[2] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, “SMOTE: Synthetic minority over-sampling technique,” J. Artif. Intell. Res., vol. 16, no. 1, pp. 321–357, Jan. 2002.
[3] F. Vilariño, P. Spyridonos, J. Virriä, and P. Radeva, “Experiments with SVM and stratified sampling with an imbalanced problem: Detection of industrial contractions,” in Proc. Int. Conf. Pattern Recognit. Image Anal., in Lecture Notes in Computer Science, vol. 3687, Aug. 2005, pp. 783–791.
[4] X. Qin, F. M. Bui, and H. H. Nguyen, “Learning from an imbalanced and limited dataset and an application to medical imaging,” in Proc. IEEE Pacific Rim Conf. Commun., Comput. Signal Process. (PACRIM), Victoria, BC, Canada, Aug. 2019, pp. 1–6.
[5] T. Karras, T. Aila, S. Laine, and J. Lehtinen, “Progressive growing of GANs for improved quality, stability, and variation,” in Proc. 6th Int. Conf. Learn. Represent. (ICLR), Vancouver, BC, Canada, Apr. 2018, pp. 1–26.
[6] Z. Lin, Z. He, S. Xie, X. Wang, J. Tan, J. Lu, and B. Tan, “AAANet: Adaptive attention network for COVID-19 detection from chest X-ray images,” IEEE Trans. Neural Netw. Learn. Syst., vol. 32, no. 11, pp. 4781–4792, Nov. 2021.
[7] D. S. Kermany et al., “Identifying medical diagnoses and treatable diseases by image-based deep learning,” Cell, vol. 172, no. 5, pp. 1122–1131, Feb. 2018.
[8] Y. Xie and D. Richmond, “Pre-training on grayscale ImageNet improves medical image classification,” in Proc. Eur. Conf. Comput. Vis. (ECCV) Workshops, Munich, Germany, Sep. 2018, pp. 476–484.
[9] A. Shamsh, H. Asharnezhad, S. S. Jokandan, A. Khosravi, P. M. Kebría, D. Nahavandi, S. Nahavandi, and D. Srinivasan, “An uncertainty-aware transfer learning-based framework for COVID-19 diagnosis,” IEEE Trans. Neural Netw. Learn. Syst., vol. 32, no. 4, pp. 1408–1417, Apr. 2021.
[10] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial nets,” in Proc. 27th Int. Conf. Neural Inf. Process. Syst., vol. 2, Montreal, QC, Canada, Jan. 2014, pp. 2672–2680.
[11] M. Arjovsky, S. Chintala, and L. Bottou, “Wasserstein generative adversarial networks,” in Proc. 34th Int. Conf. Mach. Learn., vol. 70, Sydney, NSW, Australia, Aug. 2017, pp. 298–307.

[12] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet classification with deep convolutional neural networks,” in Proc. Adv. Neural Inf. Process. Syst., vol. 2, Dec. 2012, pp. 1097–1105.

[13] S. Barratt and R. Sharma, “An ensemble of fine-tuned convolutional neural networks for medical image classification,” IEEE J. Biomed. Health Inform., vol. 21, no. 1, pp. 31–40, Jan. 2017.

[14] O. Ronneberger, P. Fischer, and T. Brox, “U-Net: Convolutional networks for biomedical image segmentation,” in Proc. Int. Conf. Med. Image Comput. Assist. Intervent., in Lecture Notes in Computer Science, vol. 9351, Oct. 2015, pp. 234–241.

[15] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Las Vegas, NV, USA, Jun. 2016, pp. 770–778.

[16] H. Li, S. Zhuang, D.-A. Li, J. Zhao, and Y. Ma, “Benign and malignant classification of mammogram images based on deep learning,” Biomed. Signal Process., vol. 51, pp. 347–354, May 2019.

[17] L. A. Jeni, J. F. Cohn, and F. De La Torre, “Facing imbalanced data—Recommendations for the use of performance metrics,” in Proc. Humaine Assoc. Conf. Affect. Comput. Intell. Interact., Geneva, Switzerland, Sep. 2013, pp. 245–251.

[18] M. Arjovsky, S. Chintala, and L. Bottou, “Towards principled methods for training generative adversarial networks,” inProc. 5th Int. Conf. Learn. Represent. (ICLR), Toulon, France, Jan. 2017, pp. 1–17.

[19] S. Lazebnik, C. Schmid, and J. Ponce, “Pyramid methods in image processing,” RCA Eng., vol. 29, no. 6, pp. 31–41, Nov. 1984.

[20] H. He and E. A. Garcia, “Learning from imbalanced data,” IEEE Trans. Knowl. Data Eng., vol. 21, no. 9, pp. 1263–1283, Sep. 2009.

[21] J. Davis and M. Goadrich, “The relationship between precision-recall and ROC curves,” in Proc. 23rd Int. Conf. Mach. Learn., New York, NY, USA, Jun. 2006, pp. 233–240.

[22] A. P. Bradley, “The use of the area under the ROC curve in the evaluation of machine learning algorithms,” Pattern Recognit., vol. 30, pp. 1145–1159, Jul. 1997.

[23] D. Davis, “Facing imbalanced data in medical imaging applications,” Expert Syst. Appl., vol. 174, Jul. 2021, Art. no. 115482.

[24] R. M. Summers, “Improving computer-aided detection using convolutional neural networks and random view aggregation,” IEEE Trans. Med. Imag., vol. 35, no. 5, pp. 1170–1181, May 2016.

[25] H. R. Roth, L. Lu, J. Liu, J. Yao, A. Seff, K. Cherry, L. Kim, and R. M. Summers, “Improved computer-aided detection using convolutional neural networks and random view aggregation,” IEEE Trans. Med. Imag., vol. 35, no. 5, pp. 1207–1226, May 2016.

[26] F. Yu, A. Seff, Y. Zhang, S. Song, T. Funkhouser, and J. Xiao, “LSUN: Construction of a large-scale image dataset using deep learning with humans in the loop,” Jun. 2016, arXiv:1506.03365, doi: 10.48550/arXiv.1506.03365.

[27] A. Krizhevsky and G. E. Hinton, “Learning multiple layers of features from tiny images,” Tech. Rep. Univ. Toronto, vol. 1, no. 4, p. 7, 2009.

[28] S. Arjovsky and L. Bottou, “Towards principled methods for training generative adversarial networks,” inProc. 5th Int. Conf. Learn. Represent. (ICLR), Toulon, France, Jan. 2017, pp. 17–18.

[29] K. Ma, Q. Wu, Z. Wang, Z. Duanmu, H. Yong, H. Li, and L. Zhang, “Group MAD competition? A new methodology to compare objective image quality models,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Las Vegas, NV, USA, Jun. 2016, pp. 1664–1673.

[30] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, “Densely connected convolutional networks,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 2261–2269.

[31] M. Tan and Q. V. Le, “EfficientNet: Rethinking model scaling for convolutional neural networks,” in Proc. 35th Int. Conf. Mach. Learn. (ICML), Jun. 2019, pp. 10691–10700.

[32] E. Yilmaz and M. Trocan, “Benign and malignant skin lesion classification comparison for three deep-learning architectures,” inProc. 12th Asian Conf. Intell. Inf. Database Syst., in Lecture Notes in Computer Science, vol. 12033, Mar. 2020, pp. 514–524.

[33] M. Jampour, S. Abbasia, and M. Javid, “CapsNet regularization and its conjugation with ResNet for signature identification,” Pattern Recognit., vol. 120, Dec. 2021, Art. no. 107851.

[34] XIAOLI QIN was born in 1991. She received the M.Eng. degree in electrical engineering from Bethel University (BUAA). She is currently pursuing the Ph.D. degree in electrical engineering with the University of Saskatchewan, Saskatoon, SK, Canada. Her research interests include machine learning and computer vision methods for medical imaging applications.

[35] FRANCIS MINHTHANG BUI (Member, IEEE) received the B.A. degree in French language and the B.Sc. degree in electrical engineering from the University of Calgary, Canada, in 2001, and the M.Sc. and Ph.D. degrees in electrical engineering from the University of Toronto, Canada, in 2003 and 2009, respectively. He is currently an Associate Professor of electrical and computer engineering with the University of Saskatchewan, Canada.
HA H. NGUYEN (Senior Member, IEEE) received the B.Eng. degree in electrical engineering from the Hanoi University of Technology (HUT), Hanoi, Vietnam, in 1995, the M.Eng. degree in electrical engineering from the Asian Institute of Technology (AIT), Bangkok, Thailand, in 1997, and the Ph.D. degree in electrical engineering from the University of Manitoba, Winnipeg, MB, Canada, in 2001.

He joined the Department of Electrical and Computer Engineering, University of Saskatchewan, Saskatoon, SK, Canada, in 2001, and became a Full Professor, in 2007. He currently holds the position of the NSERC/Cisco Industrial Research Chair in low-power wireless access for sensor networks. He is the coauthor, with Ed Shwedyk, of the textbook A First Course in Digital Communications (published by Cambridge University Press). His research interests include communication theory, wireless communications, and statistical signal processing. He is a fellow of the Engineering Institute of Canada (EIC) and a Registered Member of the Association of Professional Engineers and Geoscientists of Saskatchewan (APEGS). He was the General Chair of the 30th Biennial Symposium on Communications in 2021; the Co-Chair of the Multiple Antenna Systems and Space-Time Processing Track; IEEE Vehicular Technology Conferences, Ottawa, ON, Canada, in Fall 2010, and Quebec, QC, Canada, in Fall 2012; a Lead Co-Chair of the Wireless Access Track, IEEE Vehicular Technology Conferences, Vancouver, BC, Canada in Fall 2014; a Lead Co-Chair of the Multiple Antenna Systems and Cooperative Communications Track; IEEE Vehicular Technology Conference, Montreal, QC, in Fall 2016; and a Technical Program Co-Chair of Canadian Workshop on Information Theory, St. John’s, NL, Canada, in 2015. He was an Associate Editor of the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS and the IEEE WIRELESS COMMUNICATIONS LETTERS, during 2007–2011 and during 2011–2016, respectively. He also serves as an Associate Editor for the IEEE TRANSACTIONS ON COMMUNICATIONS and the IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY.

ZHU HAN (Fellow, IEEE) received the B.S. degree in electronic engineering from Tsinghua University, in 1997, and the M.S. and Ph.D. degrees in electrical and computer engineering from the University of Maryland, College Park, MD, USA, in 1999 and 2003, respectively. From 2000 to 2002, he was an Research and Development Engineer of JDSU, Germantown, MD. From 2003 to 2006, he was a Research Associate at the University of Maryland. From 2006 to 2008, he was an Assistant Professor at Boise State University, Boise, ID, USA. He is currently a John and Rebecca Moores Professor with the Electrical and Computer Engineering Department as well as with the Computer Science Department, University of Houston, Houston, TX, USA. His research interests include wireless resource allocation and management, wireless communications and networking, game theory, big data analysis, security, and smart grid. He has been an AAAS Fellow, since 2019, and an ACM Distinguished Member, since 2019. He received an NSF Career Award in 2010, the Fred W. Ellersick Prize of the IEEE Communication Society in 2011, the EURASIP Best Paper Award for the Journal on Advances in Signal Processing in 2015, the IEEE Leonard G. Abraham Prize in the field of Communications Systems (Best Paper Award in IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS) in 2016, and several best paper awards in IEEE conferences. He was an IEEE Communications Society Distinguished Lecturer, from 2015 to 2018. He is a 1% Highly Cited Researcher, since 2017, according to Web of Science. He is also the Winner of the 2021 IEEE Kiyo Tomiyasu Award, for outstanding early to mid-career contributions to technologies holding the promise of innovative applications, with the following citation for contributions to game theory and distributed management of autonomous communication networks.

* * *