Utilizing Generative Adversarial Network for Synthetic Image Generation to Address Imbalance Challenges in Chest X-Ray Image Classification

Nugraha Priya Utama[1,2,*] **& Muhammad Faris Muzakki**[1]

[1]School of Electrical Engineering and Informatics, Institut Teknologi Bandung,
Jalan Ganesa No. 10, Bandung 40132, Indonesia
[2]Center for Artificial Intelligence (U-COE AI-VLB), Institut Teknologi Bandung,
Jalan Ganesa No. 10 Bandung 40132, Indonesia
* E-mail: utama@staff.stei.itb.ac.id

Abstract. Deep learning-based classifiers need lots of image data to train. Unfortunately, not all real-world cases are supported by a huge amount of image data. One of the cases are images for classification of pneumonia infections with chest X-rays images. This study proposes a way of synthesizing chest X-rays with abnormal conditions in order to use the synthesized images for classification purposes. A GAN-based technique can generate synthetic images with greater quality that resemble original images thus can provide a more balanced data distribution than other approaches. To indirectly evaluate the quality of our GAN-based synthetic images, we used CNN-based classification architectures on diverse datasets. Three scenarios examined the effects of synthetic picture categorization. Scenario-1: adding 90% of synthesized images to the original images into the training dataset. Scenario-2: adding 50% of synthesized images to the original images. Scenario-3: adding 10% of synthesized image to the original images. The classification test revealed significantly increased F1 scores in all scenarios. Our study also emphasizes the significance of addressing the problem of imbalanced collections of chest X-ray images and the capability of GANs to alleviate this issue.

Keywords: *chest X-ray datasets; Covid-19; generative adversarial networks (GANs); data augmentation; imbalanced data; medical imaging; pneumonia classification; synthetic X-ray images.*

1 Introduction

The domain of medical image analysis has experienced notable progression in recent times, wherein deep learning methodologies have emerged as potent instruments for a diverse array of diagnostic and prognostic endeavors. The classification of chest X-ray images holds significant importance in the timely identification and evaluation of diverse pulmonary ailments. One of the persistent challenges encountered in this particular domain pertains to the inherent disparity observed within datasets. Specifically, there exists an imbalance between the

Received November 2nd, 2023, Revised December 13th, 2023, Accepted for publication December 31st, 2023.
Copyright © 2023 Published by IRCS-ITB, ISSN: 2337-5787, DOI: 10.5614/itbj.ict.res.appl.2023.17.3.6

quantity of positive cases, denoting instances with abnormalities, and the quantity of negative cases. It is worth noting that the former is frequently found to be considerably lower in number when compared to the latter. The presence of this disparity presents a significant hindrance to the advancement of precise and resilient machine learning models. These models have a tendency to exhibit bias in favor of the more prevalent class, which in turn may result in less than optimal clinical results.

In order to tackle the aforementioned matter, this paper delves into the pioneering methodology of employing generative adversarial networks (GANs) for the purpose of generating synthetic images within the realm of chest X-ray image classification. GANs have garnered significant attention and acclaim within the realm of computer vision due to their remarkable capacity to produce synthetic data that exhibits a striking resemblance to real data. This characteristic of GANs offers a potential solution to the prevailing issue of limited availability of positive cases. The primary objective of this research endeavor was to effectively utilize GANs in order to generate synthetic chest X-ray images that depict abnormal cases. This will ultimately facilitate the augmentation of datasets that suffer from an imbalance in the distribution of abnormal cases, thereby improving the performance of classification models.

In this paper, we will elucidate a thorough investigation into the methodology utilized for the production of synthetic chest X-ray images, the training of GANs on extensive chest X-ray datasets, and the seamless incorporation of these synthetic images into a cutting-edge classification framework. Furthermore, we assessed the influence of synthetic data augmentation on the performance, robustness, and generalization capabilities of the classification model. Our primary objective was to enhance the sensitivity and specificity for the early detection of diseases.

The salient aspect of this study resides in its pioneering utilization of GANs to tackle the inherent difficulties presented by imbalanced collections of chest X-ray images. The overarching objective was to enhance the precision and dependability of diagnostic systems employed in the realm of clinical practice. Through the mitigation of limitations imposed by imbalances within the dataset, the primary objective of this research endeavor was to offer a highly valuable solution that has the potential to greatly benefit healthcare professionals and, of utmost significance, patients. This solution aims to enhance the early detection of pulmonary conditions, thereby contributing to improved healthcare outcomes.

2 Materials And Methods

2.1 Generative Adversarial Network

A generative adversarial network is a neural network-based architecture that utilizes a discriminator and a generator simultaneously during the learning process [1]. The generator's role is to produce synthetic data that closely resembles a real dataset, while the discriminator is trained to differentiate between fake data generated by the generator and the real data [2]. The overall GAN architecture is depicted in Figure 2.

In Figure 2, the generator is trained using the real dataset, which undergoes mathematical operations through convolution layers to generate synthetic data. Convergence is achieved when the distribution produced by the generator matches the distribution of the real data. The convergence can be observed through the error values generated by the discriminator and the generator, using the following Eq. (1):

$$log(D(x)) + log(1 - D(G(z))) \tag{1}$$

Here, D represents the discriminator and G represents the generator. The generation and discrimination process is repeated until the predetermined convergence limit is reached**.** However, GAN algorithms have certain limitations in synthetic image tasks, including instability, collapse, and low resolution [3]. To address these issues, this study incorporated modifications to the loss function using Wasserstein distance, as described by Eq. (2).

$$W(P_r, P_g) = inf_{\gamma \sim \Pi(P_r, P_g)} E_{(x,y) \sim \gamma}[c(x, y)] \tag{2}$$

In this formula, W represents the Wasserstein distance, P_r represents the real distribution, P_g represents the generated distribution, and γ represents the joint distribution between the two. By using this modified loss function, the training process becomes more stable, resulting in higher image resolution and overcoming the limitations of the original GAN algorithm [4].

2.2 Dataset Description

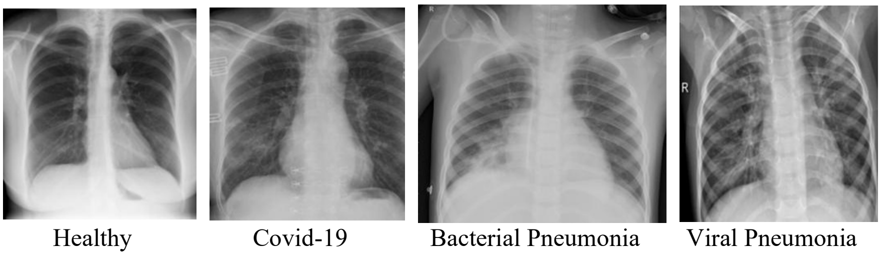
To develop an effective pre-diagnosis model, it is important to have a substantial amount of well-balanced data [5]. However, obtaining a sufficiently large and balanced dataset of X-ray images for pneumonia infections can be challenging due to factors such as cost and local privacy regulations. Therefore, this research project addressed this issue by gathering X-ray data on pneumonia infections from various sources, including Kemarny, *et al.* [6], Rahman, *et al.* [7], and Chowdhury, *et al.* [8], resulting in a total of 17,984 data samples. Detailed labels for the data can be found in Table 1.

To ensure the reliability of the collected data, a quality check was conducted on the X-ray images. Any images that were damaged or irrelevant to the study were excluded. The data collected from different sources were then combined and appropriately labeled, as indicated in Table 1. The datasets comprised multiple classes, each representing a specific type of pneumonia infection. For a visual representation of each class, example images are presented in Figure 1.

Table 1 Composition of the datasets.

Label Name	Number of Images
Healthy	10.192
Covid-19	3.621
Bacterial Pneumonia	2.826
Viral Pneumonia	1.345

The imbalanced chest X-ray datasets comprised Covid-19, healthy, bacterial pneumonia, and viral pneumonia.



Healthy Covid-19 Bacterial Pneumonia Viral Pneumonia

Figure 1 Sample data. Samples of chest X-ray images from a healthy patient and patients with Covid-19, bacterial pneumonia, and viral pneumonia.

2.3 Model Evaluation

To evaluate the performance of each research output, various metrics were utilized. The first metric employed was the Frechet Inception Distance (FID) [9], which assesses both the quality and distribution of the generated images from the generative adversarial network. In the context of medical research, FID is commonly used to measure image quality by comparing the similarity between generated and real images. It calculates the distance between the feature distributions extracted from the generated and real images using the following Eq. (3):

$$FID(P,Q) = \left\|mu_P - mu_Q\right\|^2 + Tr\left(C_P + C_Q - 2 * \left(C_P * C_Q\right)^{0.5}\right) \tag{3}$$

Here, P represents the feature distribution of the real or reference images, while Q represents the feature distribution of the generated images. u_p and u_q denote the mean feature distributions of P and Q, respectively, while C_P and C_Q represent the covariance matrices of the feature distributions of P and Q, respectively. A smaller FID value indicates a closer resemblance between the generated and the real images.

Another metric employed was the non-parametric statistical test known as Mann-Whitney U [10]. This is used to measure the distribution of the data and determine its significance. The Mann-Whitney U test compares two independent samples to ascertain whether they are drawn from the same population. The formula for calculating the Mann-Whitney U value is in Eq. (4):

$$U = R_1 - \frac{n_1 * (n_1 + 1)}{2} \tag{4}$$

In this formula, R_1 represents the sum of ranks of the first sample, while n_1 represents the number of samples in the first sample. The Mann-Whitney U test utilizes the rank of the data rather than the actual values. A threshold of 0.05 is commonly applied to determine whether a significant difference exists between the distributions of the real and the generated images.

3 Result and Discussion

3.1 Synthetic Images

The GAN generator network developed in the preceding phase was employed to generate synthetic images. Specifically, this study generated 15,000 synthetic images for each label. Following the generation process, these images were fed into a discriminator network to assess their quality. To determine whether a synthetic image should be included in the final dataset, a threshold of 0.5 was utilized, meaning that only images deemed authentic by the discriminator were selected. The distribution of the images for each label in the resulting datasets is presented in Table 2, and an example synthetic image can be observed in Figure 2.

Table 2 Synthetic image composition.

Label	Number of generated synthetic images	Number of synthetic images after selection (discriminator error > 0.5)
Covid-19	15,000	14,995
Bacterial Pneumonia	15,000	5,680
Viral Pneumonia	15,000	15,000

We produced 15,000 synthetic chest X-ray images for each category with the trained GANs. We then selected only those synthetic images which had discriminator error >0.5 to ensure the authenticity of the synthetic images for each category.

Table 2 indicates a significant decrease in the number of images, approximately three times fewer than the number of images generated by GAN, specifically in the bacterial pneumonia label. This reduction may be attributed to the suboptimal quality of the synthetic images produced by GAN. Several factors could contribute to the suboptimal GAN output, such as dataset quality, architecture, or GAN parameters. To evaluate the quality of the generated synthetic images, researchers employed the Frechet Inception Distance (FID) metric, which measures both the quality and diversity of the synthetic images. For comparison purposes, the researchers also included FID values from SMOTE conventional augmentation techniques, and conventional GANs. The FID values for each approach are presented in Table 3.

Table 3 FID scores.

Label	**FID Score**					
	Our Study	**SMOTE**	**LSGAN**	**DCGAN**	**Conventional Augmentation**	**Vanilla GAN**
Covid	4,015	7,015	16,167	72,555	28,642	494,668
Bacterial Pneumonia	2,822	7,833	52,427	89,693	32,714	469,939
Viral Pneumonia	5,919	11,787	47,251	43,127	36,806	502,041

Evaluation of the quality of the synthetic images generated by our study and alternative approaches. Lower FID-score indicates more resemblance of the generated synthetic image to original images.

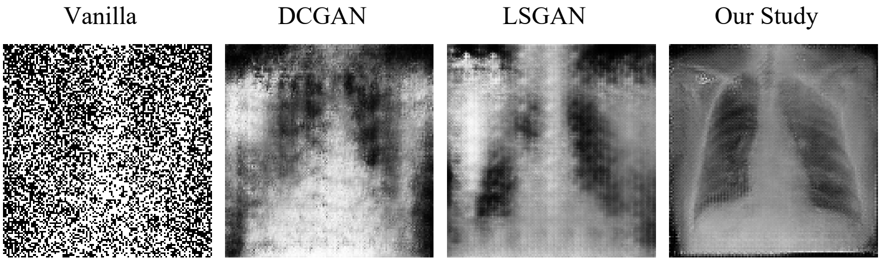
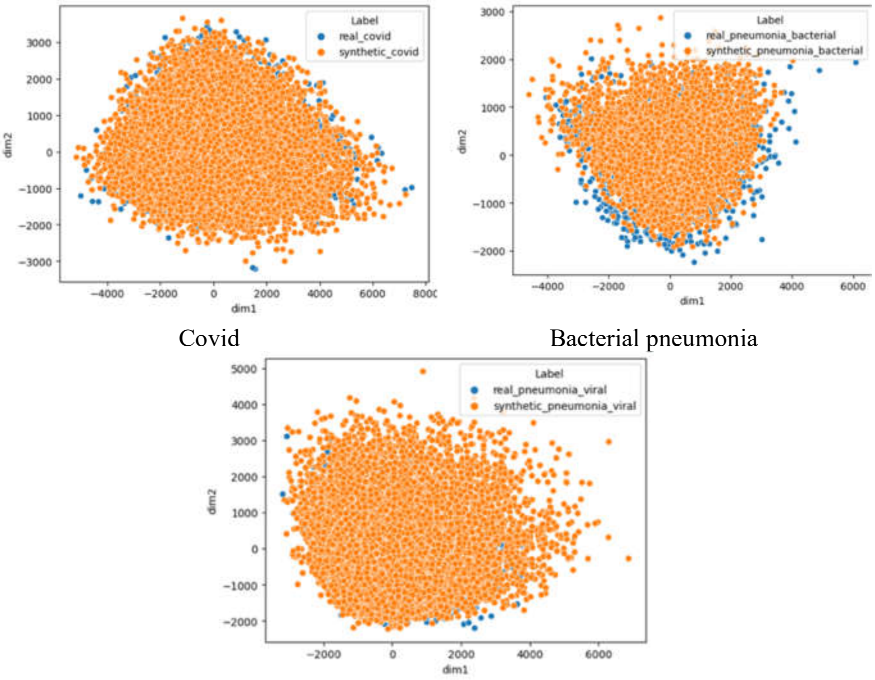


Figure 2 Image synthesis. Sample of generated synthetic images using several alternative approaches.

According to Table 3 and Figure 2, the FID value for GAN was lower compared to the other techniques. This finding suggests that the synthetic data generated by our study exhibits better image quality, closely resembling original images, and demonstrates a more balanced diversity in the data distribution compared to the other algorithms. The distribution of the synthetic dataset generated by GAN was visualized using a two-dimensional PCA algorithm, as depicted in Figure 3. Figure 3 provides a comparison of the original and the synthetic images at ratios of 3,621:14,995, 2,826:5,680, and 1,345:15,000, respectively.

The distribution results of the GAN synthetic data in Figure 3 demonstrate that the image synthesis did not always produce identical or identical images, nor did it deviate significantly from the distribution of the original images. However, in Figure 4, the synthetic data for viral pneumonia shows some data points that deviate from the central distribution of the original data. This can be attributed to the limited availability of labeled viral pneumonia datasets, which hampers the formation of a robust GAN model compared to other labels.

In addition to analyzing the per-instance data distribution, the researchers also conducted a data analysis in the Gaussian distribution domain. Figure 3 reveals that the distribution pattern of the synthetic data still followed a similar spread as the original data, albeit with relatively wide differences in mean and standard deviation values in some cases. However, the synthetic data tended to approach a normal distribution, which indicates a correct pattern for the number of data .



Covid

Bacterial pneumonia

Viral pneumonia

Figure 3 Data distribution of chest X-ray images for each category. The distribution pattern of the synthetic data (colored orange) follows a similar spread as the original data (colored blue), albeit with relatively wide differences in mean and standard deviation values in some cases. However, the synthetic data tends to approach a normal distribution.

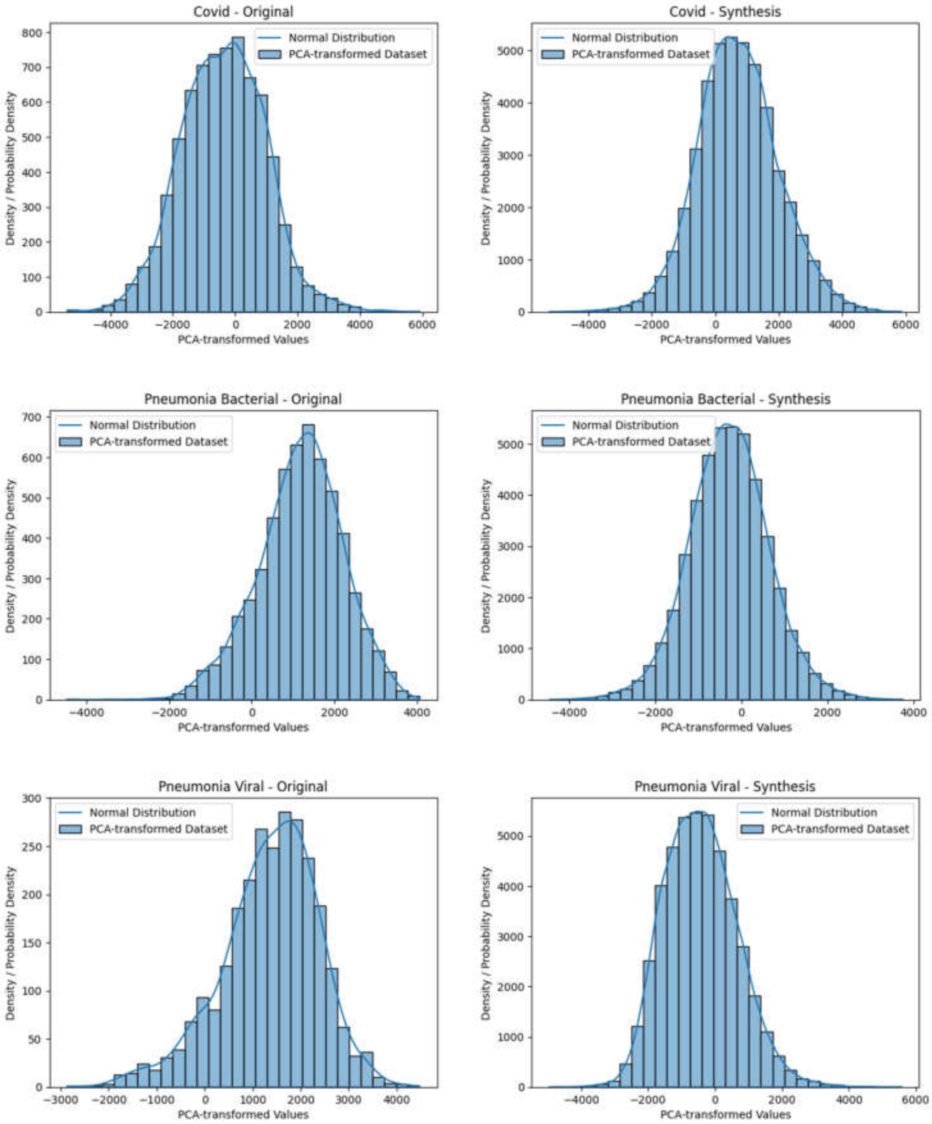


Figure 4 Data distribution of original and synthetic images on a Gaussian chart. Comparison of the original data distribution (left) and the synthetic data distribution (right).

4 Classification

A classification model was utilized to evaluate the results of GAN's synthetic image generation. In this study, we employed CNN-based architectures commonly used in image classification research to assess the performance of the classification models. The F1 scores for each classification model, with testing-to-training data ratios of 10:90, 50:50, and 90:10, respectively, running for 50 epochs, are presented in Table 4.

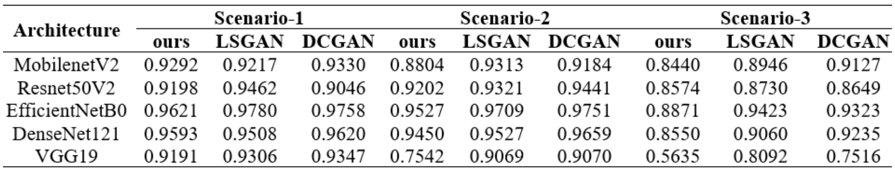
From Table 4, it can be seen that GAN-generated data could improve the F1 score across all architectures for 10% and 50% original training data. However, there was a decrease in the F1 score for the EfficientNetB0 and DenseNet121 architectures when using 90% original training data. To determine the statistical significance of the improvement in the F1 score resulting from the addition of GAN-generated training data, the researchers conducted a Mann-Whitney U nonparametric statistical test with a threshold p-value of 0.05 [11-10].

Table 4 Classification result – additional synthetic images to the datasets under three scenarios for classification tasks.

Architecture	Scenario-1		Scenario-2		Scenario-3	
	all ori	**+90% synth**	**all ori**	**+50% synth**	**all ori**	**+10% synth**
MobilenetV2	0.6752	0.9454	0.7817	0.9348	0.4138	0.9049
Resnet50V2	0.9175	0.9330	0.9110	0.9121	0.8113	0.8870
EfficientNetB0	0.9797	0.9780	0.9660	0.9713	0.8863	0.9446
DenseNet121	0.9612	0.9484	0.9506	0.9517	0.8492	0.9029
VGG19	0.9190	0.9294	0.8717	0.9174	0.6708	0.8529

In this study, a p-value of 0.03147 was obtained based on the data from Table 4. This result indicates that the addition of GAN-generated synthetic data provides a significant improvement in the F1 score for the classification results based on the conducted testing scenarios. In addition to comparing the F1 scores with non-synthetic data models, the authors also compared the classification outcomes with GAN algorithms from previous studies. The F1 scores for each approach can be found in Table 5**.**

Table 5 Comparison of our synthetic images with other synthetic images from different classification techniques.



Architecture	Scenario-1			Scenario-2			Scenario-3		
	ours	**LSGAN**	**DCGAN**	**ours**	**LSGAN**	**DCGAN**	**ours**	**LSGAN**	**DCGAN**
MobilenetV2	0.9292	0.9217	0.9330	0.8804	0.9313	0.9184	0.8440	0.8946	0.9127
Resnet50V2	0.9198	0.9462	0.9046	0.9202	0.9321	0.9441	0.8574	0.8730	0.8649
EfficientNetB0	0.9621	0.9780	0.9758	0.9527	0.9709	0.9751	0.8871	0.9423	0.9323
DenseNet121	0.9593	0.9508	0.9620	0.9450	0.9527	0.9659	0.8550	0.9060	0.9235
VGG19	0.9191	0.9306	0.9347	0.7542	0.9069	0.9070	0.5635	0.8092	0.7516

Table 5 shows that our generated synthetic data had comparable results for classification tasks, especially in Scenario-1 and Scenario-2. However, our generated synthetic data had lower FID scores, which indicates resemblance to the original images. This phenomenon indicates that for classification tasks with deep learning, the higher semantic information in the synthetic images is not significantly important, but it is important for human subjects. This indicates the difference in nature of the processes in our brain and the process of classification with deep learning.

5 Conclusion

In this study, data synthesis using GAN algorithms was conducted to address the imbalanced data issue in pneumonia infection classification using X-ray image data as input. The findings of this study indicate that the constructed GAN architecture produced synthetic data that closely approximates the quality of original data, while exhibiting a more diverse data distribution compared to conventional augmentation algorithms. The GAN architecture developed in this study also demonstrated better resolution and relative resistance to collapse compared to the tested GAN algorithms. Moreover, the synthetic data utilized in the construction of the classification model significantly enhanced the F1 score based on the results of statistical tests.

For future research, several improvements can be made, including the utilization of more advanced GAN architectures such as Big GAN and Style GAN. The incorporation of these GAN variants may enhance image sharpness at high resolutions. In addition to the basic GAN architectures, employing a penalty system instead of gradient clipping may provide more stable image quality. Involving experts, specifically radiologists, in the research would have a significant impact on expanding the scope of analyses that can be performed. In this study, quantitative analysis was only conducted on the overall image quality. However, by involving experts in the quantitative analysis process from a medical perspective, there is a greater possibility of discovering new findings and ideas.

References

[1] Goodfellow, I.J., Pouget-Abadie J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. & Bengio, Y., *Generative Adversarial Networks,* arXiv preprint arXiv:14062661, 2014.

[2] Mahdizadehaghdam, S., Panahi, A. & Krim, H., *Sparse Generative Adversarial Network,* Proceedings - 2019 International Conference on Computer Vision Workshop, ICCVW 2019, pp. 3063-3071. DOI:10.1109/ICCVW.2019.00369, 2019.

[3] Mirza, M. & Osindero, S., *Conditional Generative Adversarial Net,* 1-7. Available: http://arxiv.org/abs/1411.1784, 2014.

[4] Arjovsky, M., Chintala, S. & Bottou L., *Wasserstein Generative Adversarial Networks*, in Precup, D. & The, Y.W. (Eds.) Proceedings of the 34th International Conference on Machine Learning, PMLR., pp. 214-223, 2017.

[5] Santosh, K.C., *AI-Driven Tools for Coronavirus Outbreak: Need of Active Learning and Cross-Population Train/Test Models on Multitudinal/Multimodal Data*, J Med Syst., **44**, 93, 2020. DOI:10.1007/s10916-020-01562-1.

[6] Kermany, D.S., Goldbaum, M., Cai, W. & Lewis, M.A., *Identifying Medical Diagnoses and Treatable Diseases by Image-Based Deep Learning Resource Identifying Medical Diagnoses and Treatable Diseases by Image-Based Deep Learning,* Cell, **172**, pp. 1122-1131.e9, 2018. DOI: 10.1016/j.cell.2018.02.010.

[7] Rahman, T., Khandakar, A., Qiblawey, Y., Tahir, A., Kiranyaz, S., Abul Kashem, S., Bin, et al. *Exploring the Effect of Image Enhancement Techniques on COVID-19 Detection using Chest X-Ray Images,* Comput Biol Med, **132**, 104319, 2021. DOI: 10.1016/j.compbiomed.2021.104319,

[8] Chowdhury, M.E.H., Rahman, T., Khandakar, A., Mazhar, R., Kadir, M.A., Mahbub, Z. Bin, Islam, K.R., Khan, M.S., Iqbal, A., Al Emadi, N., Reaz, M.I. & Islam, M.T., *Can AI Help in Screening Viral and COVID-19 Pneumonia?,* IEEE Access, **8**, pp.132665-132676, 2020. DOI:10.1109/ACCESS.2020.3010287.

[9] Brock, A., Donahue, J. & Simonyan, K., *Large Scale GAN Training for High Fidelity Natural Image Synthesis,* 2018.

[10] Perme, M.P. & Manevski, D., *Confidence Intervals for the Mann-Whitney Test,* Stat Methods Med Res, **28**, pp. 3755-3768, 2019. DOI:10.1177/0962280218814556,

[11] Cho, Y.K. & Kim, M.S., *Dry Eye After Cataract Surgery and Associated Intraoperative Risk Factors,* Korean Journal of Ophthalmology, **23**, 65, 2009. DOI:10.3341/kjo.2009.23.2.65.