Enhancing Medical Image Classification Through GAN-Augmented Datasets: A Comparative Study on Cardiomegaly and Pneumonia Detection

Morgan Benavidez [1]
Department of Electrical Engineering and Computer Science, Florida Atlantic University
Boca Raton, Florida

***Abstract*—In deep learning and computer vision, the effectiveness of models heavily depends on the availability and diversity of training data. However, in the medical field, acquiring a sufficient amount of labeled data for model training is challenging due to privacy laws. This report explores an approach to mitigate the effects of limited training data by using Generative Adversarial Networks (GANs). In this report I demonstrate how GANs can be used for generating synthetic images to augment and increase the training dataset for deep learning models. By training a GAN on existing data and generating synthetic images, I aim to strengthen the training set, making it more robust and comprehensive. My research not only explores the technical aspects of GAN-based data generation but also examines the implications of using synthetic data on the performance of deep learning models. The outcomes of this study provide valuable insights into the feasibility and effectiveness of using GAN-generated synthetic data to improve the performance of deep learning models with medical data. This research bridges the gap between limited data availability and the demands of modern machine learning applications, offering a promising avenue for addressing real-world challenges in various domains outside of medicine. By leveraging this framework, this project contributes to the ongoing efforts to advance the field of machine learning, and lays the groundwork for applications in other fields where data scarcity has been a bottleneck.**

I. Introduction

A. Define the Problem

The challenge of limited training data in medical image analysis is a critical bottleneck, due to HiPPA laws protecting patients's privacy. The issue of limited data is additionally compounded for diseases that are not as common and have fewer samples [6], [7], [8], [13]. This scarcity impedes the development of robust diagnostic models, but also reflects a broader issue in healthcare: the need for accurate, AI-driven medical diagnoses. Implementing AI in healthcare will likely lead to significantly improvevements in correctly diagnosing diseases, as well as the early detection of diseases. This implementation is critical, as the advancements will allow for better treatement of the disease and symptoms, leading to an improved quality of life for patients. This study seeks to address the limitation of integrating AI into healthcare by using GANs to generate synthetic medical training data [1], [3], [5]-[8], [12]-[15]. The goal of this procedure is to provide a framework that can be used to improve the diagnostic accuracy of medical imaging in resource-constrained scenarios.

B. Motivation

The motivation for this research is the urgent need for improved medical diagnostics through AI. With the growing prevalence of diseases and increasing demands on healthcare systems, efficient and accurate diagnostic tools are essential. AI offers significant promise in revolutionizing healthcare, particularly in medical imaging, by providing faster and more accurate diagnoses. The better identification and treatment of diseases will likely lead to better patient care and overall outcomes. Additionally, AI can potentially reduce physician workload and healthcare costs, making quality medical services more accessible. This study focuses on how using GANS can address the challenge of limited data in diagnosing diseases from chest X-rays, such as pneumonia and cardiomegaly. This project serves as a foundation, demonstrating the effectiveness of this technique in a small subset of data. The intention is that this framework can be adapted to use in a variety of settings where data is sparse, and might include applications outside of the healthcare field [5], [6], [8], [11], [12].

However, these conventional augmentation methods have limitations, particularly in the context of medical imaging. While they can increase dataset size and variability, they may fall short in generating clinically relevant variations. This is especially true for rare diseases or subtle pathological features that require nuanced representation in training datasets. Such limitations highlight the need for more advanced augmentation techniques and generative models, like Generative Adversarial Networks (GANs), which can create new, realistic images that capture the complexity and variability inherent in medical imaging. GANs, for instance, can generate images with pathologies or conditions not present in the original dataset, thereby enriching the training data with a wider range of clinical scenarios [13].

C. Overview of Methods

Generative Adversarial Networks refer to a class of algorithms used in unsupervised machine learning, which are implemented by a system of two neural networks contesting with each other. One network, the generator, creates data that is as realistic as possible from the training data. The other network is called the discriminator, and it evaluates the data for authenticity. The process continues until the generated

data is indistinguishable from real data to the discriminator. This method makes GANs particularly effective for generating complex and diverse data, such as medical images, where real-world data may be scarce or sensitive [5].

GANs offer significant advantages over traditional data augmentation techniques in medical image classification. While traditional augmentation can only modify existing images through techniques like rotation, cropping, or flipping, GANs can create entirely new images that maintain anatomical correctness and variability. This results in a more realistic images, producing an extensive and varied dataset. By generating high-quality, diverse synthetic images, GANs can greatly enhance the depth and breadth of training datasets. This practice is particularly useful for AI applications, since deep learning models often require a large number of images for training [10], [11].

Specifically, this project focuses on enhancing medical image classification through a tailored CNN for detecting specific conditions like pneumonia and cardiomegaly from chest X-rays. CNNs were chosen because they are particularly good at feature extraction and pattern recognition in image data. This model incorporates several layers including convolutional layers with diverse filter sizes for detailed feature detection, max pooling layers for dimensionality reduction, and dropout layers to prevent overfitting. To counteract the challenge of limited data, I have integrated GANs for data augmentation, aiming to enhance the training dataset with synthetic, yet realistic images. This comprehensive approach seeks to improve the accuracy and effectiveness of medical image classification, demonstrating the potential impact of AI in medical diagnostics under data-constrained situations.

II. Proposed Method

A. General Workflow

This study employs a dual approach, integrating Convolutional Neural Networks and Generative Adversarial Networks for medical image analysis and generation. CNNs are ideal models for extracting features from medical images, and will be used with samples from the MedMNIST dataset. The GAN will be trained on the MedMNIST images to create realistic synthetic samples. Next, I have created a Structural Similarity Index (SSI) for comparative analysis between real and generated images, ensuring that the GAN has created images that are realistic and similar to the original training data. This methodology provides a robust framework for enhancing medical image analysis and generation, aiming to improve both accuracy and applicability in healthcare research [2].

B. Related Work: Existing Methods for Augmenting Data

In addressing the issue of limited data availability in medical imaging, current AI methods primarily focus on advanced data analysis and classical augmentation techniques.The premise behind these methods is to maximize the utility of existing datasets and enhance the diversity and volume of training data, which is crucial for the effective training of deep learning models. Traditional data augmentation methods include rotating, flipping, or scaling images, which is used to increase the diversity of datasets. These methods serve to artificially expand the dataset by creating variations of the existing images, simulating different viewing angles, orientations, and sizes. This diversity is critical in training robust models capable of accurately analyzing medical images under various conditions. By exposing the model to a broader range of data representations, augmentation helps in mitigating overfitting, where a model performs well on training data but poorly on unseen data [4], [10].

C. Related Work: Common approaches to transfering knowledge in image analysis

In this project, we demonstrate the utilization of Generative Adversarial Networks (GANs) to augment a medical imaging training dataset, subsequently applying the enhanced dataset to a different type of medical scan as a baseline. Central to this approach is the concept of transfer learning, which is instrumental in enabling the generalization of results from one type of medical imaging to another. Transfer learning allows the model trained on the GAN-augmented dataset to adapt its learned features to a new, yet related, dataset. This adaptability is crucial in medical imaging, where different types of scans may share underlying patterns and features, despite differences in their visual presentation. By leveraging transfer learning, the model not only benefits from the enriched diversity of the augmented dataset but also acquires the versatility to effectively interpret various forms of medical imagery, enhancing its diagnostic applicability across different imaging modalities [14], [15].

A common approach to transfering knowledge involves using a pre-trained model. This process involves selecting a well-established model, such as Alexnet or ResNet, and fine-tuning it for a specific, smaller dataset within the intended domain. This enables the model to become familiar with the nuances of the target domain, leading to improved performance in testing scenarios. There are several advantages to using this method. Predominantly, a large, pre-existing model encompasses a wealth of fundamental features embedded in its weights and biases. This ingrained 'background knowledge' about the task's structure is proven to not only elevate accuracy in the target domain but also diminish both the training duration and associated costs. Aditionally, this foundational knowledge streamlines the training process for new models, rendering the approach more to be less computationally expensive and more time-efficient [3], [6]-[8].

Domain adaptation is another popular technique for knowledge transfer from one domain to distinct target domains. This approach involves maintaining the same fundamental task, such as image classification or disease detection, while adjusting the model to account for the variations between the source and target domains. These variations could be in the form of different imaging techniques, varying patient demographics, or distinct disease presentations. Essentially, the method adapts the model to understand and process the

unique features of the target domain while still performing the original task it was designed for. The primary objective is to align the source domain's distribution more closely with that of the target domain. Such alignment significantly enhances the model's capability to generalize effectively across various domains, thereby boosting its accuracy in classification tasks. This technique could be especially beneficial in the field of medical imaging, since it allows models trained on one disease to improve diagnostic accuracy for similar conditions [8], [12].

III. Main Body

A. Design, Logic and Motivation

Design: GANs CNNs, Real / Fake Comparison This project utilizes a simple neural network to classify Chest X-Ray images from the MedMNIST dataset [9]. The CNN model, as detailed in the neuralNetwork.py file, is the cornerstone of this image classification system. Its architecture is designed to extract intricate features from medical images, comprising multiple convolutional layers, activation functions, and pooling layers. This design is tailored specifically for the complexities of medical images in the MedMNIST dataset, and focuses on key diagnostic features.

The GAN, outlined in the gan4.py file, plays a critical role in generating synthetic medical images. This GAN is trained on real images from the MedMNIST dataset [9], enabling it to produce images that closely resemble genuine medical scans. The generated images serve a dual purpose: augmenting the training dataset for the CNN and providing a means to test the CNN's ability to differentiate between real and synthetic images.

The project also incorporates a comparative analysis module, as seen in the compareRealFake.py file. This module uses a Structural Similarity Index (SSI) to quantitatively evaluate the similarity between real and GAN-generated images. This comparison is crucial for assessing the authenticity of the generated images and the effectiveness of the CNN in classifying them.

Greater Than Likeness	Total Generated Images	Cardiomegaly Real Images	Total Images in Training Set	Total Images After Balancing with Normal Chest-X-rays	Test Set (Balanced)	Validation Set (Balanced)
80	11415	1950	13365	26730	1164	480
81	9103	1950	11053	22106	1164	480
82	6850	1950	8800	17600	1164	480
83	4612	1950	6562	13124	1164	480
84	2906	1950	4856	9712	1164	480
85	1588	1950	3538	7076	1164	480
86	701	1950	2651	5302	1164	480
87	240	1950	2190	4380	1164	480
88	66	1950	2016	4032	1164	480
89	17	1950	1967	3934	1164	480

Fig. 1. Cardiomegaly Dataset Dimensions - These are the starting dimensions to train the GAN with and also the number of augmented images produced and their likeness ratios.

Datasets	Pneumonia Images	Pneumonia Images After Balancing with Normal Chest X-rays
Training	3494	6988
Validation	389	778
Testing	390	780

Fig. 2. Pneumonia Dataset Dimensions - These are the dimensions for the baseline dataset.

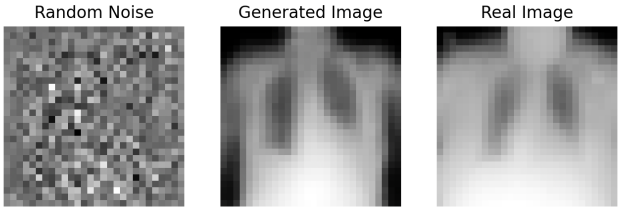

Fig. 3. The GAN starts with random noise and produces images similar to the real images.

Logic and Motivation: The motivation behind combining CNNs and GANs stems from the need to improve the accuracy and reliability of medical image classification. In medical imaging, where data can be scarce and diverse, the ability to generate additional training data and rigorously test classification models is invaluable. I suspect that the integration of GAN-generated images into the training process will enhance the CNN's performance, particularly in terms of generalization and robustness.

Furthermore, the project's focus on comparative methods to evaluate real and synthetic images is driven by the necessity to maintain high standards of accuracy in medical diagnostics. The ability to reliably distinguish between real and synthetic images ensures the integrity of the classification model in practical applications.

B. Justification for Novel Components

The proposed experimental design differs from existing work in that it is an ensemble of various methods. This model takes concepts from many areas used in old and modern machine learning. The combination of these tools should be useful in improving classification accuracy without increasing the complexity and computation cost associated with deep learning models, which is especially helpful since we have limited examples.

IV. Experiments

A. Experimental Settings

1) Programming tools and languages: The implementation of the algorithms relies on Python 3 as the primary programming language. TensorFlow serves as the key framework for developing and training both the GAN model and the CNN. Specifically, TensorFlow is employed to fine-tune the parameters of the GAN model, aiming to improve the quality of synthetic Cardiomegaly images generated by the model. The

neural network architecture remains consistent throughout the experiments. The sci-kit learn library is utilized to assess the likeness of the generated images to real images, a crucial step in the data validation process. Additionally, the NumPy library is utilized for transforming data inputs into matrices, a prerequisite for efficient manipulation of data with machine learning models. Lastly, the Matplotlib library is used for creating plots to facilitate visualizations throughout the analysis.

2) Parameters: Testing was performed using a likeness threshold of 80%, 81%, 82%, 83%, 84%, 85%, 86%, 87%, 88%, and 89% or greater was employed to curate the training dataset for Cardiomegaly, encompassing both real and augmented images generated by the GAN. The CNN training configuration utilized a batch size of 32, ensuring efficient gradient updates during optimization. K-fold cross-validation consisting of n_splits = 3 was used to optimize performance, generalize to unseen data and reduce risk of overfitting. A consistent training duration of 40 epochs was applied to the CNNs using training sets consisting of only real data across all iterations. This includes the Cardiomegaly and Pneumonia datasets. Augmented training sets greater than 20,000 used 200 epochs per k-fold, greater than 10,000 used 100 epochs per k-fold, greater than 5,000 used 80 epochs per k-fold and all others used 40 epochs per k-fold. To ensure fair comparisons, the testing set remained unchanged for both the real Cardiomegaly training sets and those augmented with synthetic images. This standardized approach aimed to systematically assess the impact of the augmented dataset on CNN performance across various training conditions.

3) Benchmark data and adding dimensions: Two benchmark datasets are utilized in this study: PneumoniaMNIST and ChestMNIST. The PneumoniaMNIST dataset serves as a benchmark for neural network performance, while the Cardiomegaly images are extracted from the ChestMNIST dataset. The Cardiomegaly dataset is smaller in size compared to PneumoniaMNIST. The shape of every image used in the study is (28, 28, 1), including all augmented data. The size and dimensions of the Cardiomegaly dataset are detailed in Figures 3 and 4, providing context for the experimental setup. The model will be trained in each iteration along with the augmented datasets as a constant to observe the behavior of the augmented training in relationship to it.

The ChestMNIST dataset contains chest X-ray images with various conditions, and the Cardiomegaly subset is identified and isolated for targeted augmentation. The size and dimensions of the Cardiomegaly dataset are detailed, providing context for the experimental setup.

4) Baseline Methods: The baseline method involves training the neural network solely with real Cardiomegaly images, establishing a standard for comparison. Simultaneously, the proposed methodology seeks to surpass this baseline by incorporating GAN-generated images into the training process. The PneumoniaMNIST dataset is used to highlight the capability of the CNN used in the training of the smaller Cardiomegaly dataset and all subsequent augmented Cardiomegaly datasets.

This comparison aims to determine if augmenting the dataset with Generative Adversarial Network (GAN)-generated images enhances the neural network's ability to generalize and accurately identify Cardiomegaly. Additionally, the inclusion of Pneumonia training serves to validate the functionality of the convolutional neural network (CNN) used for classification, offering a broader context for assessing the proposed augmentation approach.

Group	Real Training Accuracy	Gen Training Accuracy	Pneumonia Training Accuracy	Real Validation Accuracy	Gen Validation Accuracy	Pneumonia Validation Accuracy	Real Test Accuracy	Gen Test Accuracy	Pneumonia Test Accuracy
80	87.614%	97.398%	99.008%	82.192%	93.105%	97.798%	77.577%	67.955%	89.359%
81	88.596%	97.173%	98.307%	82.237%	93.009%	97.386%	76.976%	72.423%	88.333%
82	86.632%	95.708%	98.326%	81.119%	91.167%	97.554%	77.320%	71.907%	90.256%
83	87.146%	96.567%	98.957%	81.393%	90.253%	97.657%	75.945%	72.165%	90.641%
84	90.034%	94.363%	99.298%	83.219%	87.716%	97.541%	77.062%	73.024%	88.846%
85	90.662%	92.840%	98.487%	83.493%	85.946%	97.309%	75.687%	75.601%	90.513%
86	88.356%	92.744%	98.526%	82.192%	83.882%	97.747%	76.976%	76.460%	90.128%
87	85.765%	84.558%	99.066%	80.753%	79.609%	97.850%	76.460%	77.663%	89.744%
88	85.308%	89.816%	98.474%	80.342%	83.156%	97.373%	78.351%	78.007%	88.974%

Fig. 4. Experiment Results

B. Results

Figures 4-7 are the experiment results and visual representation of the models' training, validation and testing accuracies. The blue lines represent the data that included no augmented images. The orange line represents the datasets that included both real and augmented images.

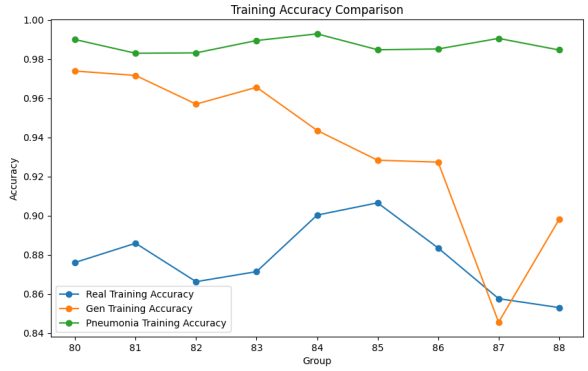

Fig. 5. Training Accuracy

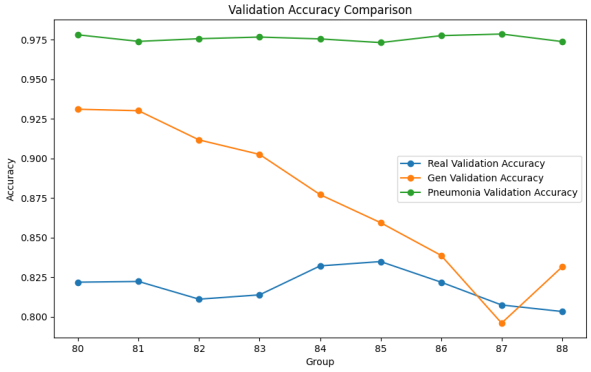

Fig. 6. Validation Accuracy

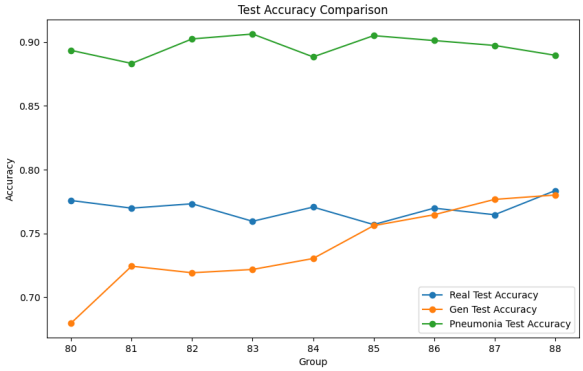

Fig. 7. Testing Accuracy

C. Analysis

As you can see in Figures 4, 5 and 6, the Pneumonia and real Cardiomegaly data tends to be more constant. This is because all of the training data for these two datasets actual images of X-rays from patients with Pneumonia. The Pneumonia has higher accuracy because the training, validation and testing sets have more images to use and can therefore yield more accurate results. The augmented Cardimegaly training and validation accuracies lower as the image likeness raises becasue there are less of these higher quality augmented images in the dataset. They are also have mostly less than a 90% likeness ratio.

What's interesting is that as the likeness ratio increases, the testing accuracy increases. This shows that as more similar augmented images are added to the data set, the better the model trained on augmented data is at classifying unseen data correctly. While the proposed approach does not perform better than the Pneumonia baseline model or Cardiomegaly model, the testing accuracy is trending towards the baseline as augmented images with a higher likeness ratio are added to the Cardiomegaly dataset.

It is also possible that Cardiomegaly is simply harder to detect from Chest X-rays than Pneumonia, or there is not enough variation in the Cardiomegaly dataset to begin with.

V. Conclusion

This experiment integrated Generative Adversarial Networks (GANs) with Convolutional Neural Networks (CNNs) for medical image classification. GANs proved effective in addressing limited and diverse datasets, offering advantages over traditional augmentation techniques. While focusing on pneumonia and cardiomegaly detection from chest X-rays, our CNN model, incorporating diverse layers, showcased the potential of GANs for data augmentation.

The study demonstrated the impact of GAN-generated synthetic images on improving model generalization. Despite not surpassing baseline models, trends in testing accuracy indicated positive outcomes with higher-likeness augmented images. Challenges in detecting cardiomegaly or limited dataset variation were acknowledged. For future investigations, it is recommended to direct attention toward the refinement of the GAN model, aiming to generate synthetic images with likeness ratios exceeding 90%. This has the potential to enhance the performance of the classification model and yield improved results.

Our work contributes insights into GAN integration for medical image classification, emphasizing potential refinements for future exploration. This dual approach holds promise for addressing data constraints in medical diagnostics, showcasing the evolving role of AI in enhancing accuracy and effectiveness.

References

[1] A. Mikołajczyk, S. Majchrowska, S. Limeros, *The (de)Biasing Effect of GAN-Based Augmentation Methods on Skin Lesion Images*, Ithaca: Cornell University Library, arXiv.org, 2022

[2] A. Salazar, L. Vergara, G. Safont, "Generative Adversarial Networks and Markov Random Fields for Oversampling Very Small Training Sets," *Expert Systems with Applications*, vol. 163, 2021. Available: https://doi.org/10.1016/j.eswa.2020.113819.

[3] C. Kwon, S. Park, S. Ko, J. Ahn, "Increasing Prediction Accuracy of Pathogenic Staging by Sample Augmentation with a GAN," *PloS one*, vol. 16, no. 4, 2021. Available: https://doi.org/10.1371/journal.pone.0250458

[4] C. Shorten, T. Khoshgoftaar, "A Survey on Image Data Augmentation for Deep Learning," *Journal of big data* , vol. 6, no. 1, 2019. pp. 1-48 Available: http://dx.doi.org/10.1186/s40537-019-0197-0

[5] E. Strelcenia and S. Prakoonwit, "Improving Cancer Detection Classification Performance Using GANs in Breast Cancer Data," *IEEE*, vol. 11, 2023, pp. 71594-71615.

[6] H. Ali, Z. Shah, "Combating COVID-19 Using Generative Adversarial Networks and Artificial Intelligence for Medical Images: Scoping Review," *JMIR Medical Informatics*, vol. 10, no. 6, 2022. Available: https://doi.org/10.2196/37365

[7] I. Amin, S. Hassan, J. Jaafar, "Semi-Supervised Learning for limited medical data using Generative Adversarial Network and Transfer Learning," in *2020 International Conference on Computational Intelligence (ICCI)*, Bandar Seri Iskandar, Malaysia, 2020, pp. 5-10.

[8] J. Mendes, T. Pereira, F. Silva, J. Frade, J. Morgado, C. Freitas, E. Negrão, "Lung CT Image Synthesis Using GANs," *Expert systems with applications*, vol. 215, 2023. Available: https://doi.org/10.1016/j.eswa.2022.119350

[9] J. Yang, R. Shi, D. Wei, Z. Liu, L. Zhao, B. Ke, H. Pfister, B. Ni, "MedMNIST v2 - A large-scale lightweight benchmark for 2D and 3D biomedical image classification," *Sci Data*, vol. 10, no. 41, 2023. Available: https://doi.org/10.1038/s41597-022-01721-8

[10] N. Waqas, S. Safie, K. Kadir, S. Khan, M. Kehl, "DEEPFAKE Image Synthesis for Data Augmentation," *IEEE Access*, vol. 10, 2022. Available: https://doi.org/10.1109/ACCESS.2022.3193668

[11] R. Toda, A. Teramoto, M. Kondo, K. Imaizumi, K. Saito, H. Fujita, "Lung Cancer CT Image Generation from a Free-Form Sketch Using Style-Based Pix2pix for Data Augmentation," *Scientific reports*, vol. 12, no. 1, 2022. Available: https://doi.org/10.1038/s41598-022-16861-5

[12] T. Pang, J. Wong, W. Ng, C. Chan, "Semi-supervised GAN-based Radiomics Model for Data Augmentation in Breast Ultrasound Mass Classification," *Computer Methods and Programs in Biomedicine*, vol. 203, 2021.

[13] X. Yi, E. Walia, P. Babyn, "Generative Adversarial Network in Medical Imaging: A Review," *Medical Image Analysis*, vol. 58, 2019. Available: https://doi.org/10.1016/j.media.2019.101552

[14] Z. Liu, Q. Lv, C. Lee, L. Shen, "GSDA: Generative Adversarial Network-Based Semi-Supervised Data Augmentation for Ultrasound Image Classification," *Heliyon*, vol. 9, no. 9, 2023. Available: https://doi.org/10.1016/j.heliyon.2023.e19585

[15] Z. Qin, Z. Liu, P. Zhu, Y. Xue, "A GAN-Based Image Synthesis Method for Skin Lesion Classification," *Computer Methods and Programs in Biomedicine*, vol. 195, 2020. Available: https://doi.org/10.1016/j.cmpb.2020.105568

[16] M. Benavidez, "HealthGAN," GitHub, 2023. [Online]. Available: https://github.com/morganbenavidez/HealthGAN