Evaluating Auxiliary GANs for Realistic Chest X-ray Synthesis: A VGG-16 Analysis in Healthcare

Aaryamonvikram Singh
Thadomal Shahani Engineering College

Riz Lala
Thadomal Shahani Engineering College

Sourav Macwan
Thadomal Shahani Engineering College

Vanshika Chaurasia
Thadomal Shahani Engineering College

ABSTRACT
Generative adversarial networks (GANs) show promise for synthesizing realistic medical images; however, the quantitative evaluation of accuracy remains difficult. This study utilizes an auxiliary GAN to generate synthetic chest X-rays and then use these images to train a VGG-16 convolutional neural network (CNN) classifier. The CNN's performance in classifying real X-rays was evaluated to assess the efficacy of GAN-generated training data. The auxiliary GAN was trained on real X-rays and then used to synthesize images modeled after the original data distribution. The VGG-16 model was trained on a synthetic dataset and tested on reserved real X-rays, which had not been seen during model development. Its performance was compared with classifiers trained solely on real data. The results analyze the VGG-16 testing accuracy between synthetic and real training data to quantify how effectively the auxiliary GAN captured the visual features critical for high CNN performance. Techniques for evaluating GAN-generated content as part of the clinical adoption of generative models are discussed. This study presents a methodology for assessing GANs in the production of synthetic medical training data while preserving vital information for analysis.

Keywords
Auxiliary Generative Adversarial Networks, Classification, Deep Learning, Neural Networks

1. INTRODUCTION
Generative adversarial networks (GANs) are an encouraging approach for generating synthetic medical images that can serve as training data for downstream applications. In particular, auxiliary GANs have been shown to be effective in synthesizing realistic chest X-rays. However, a rigorous evaluation of the accuracy of GAN-derived synthetic data remains a key challenge. This study employs an auxiliary GAN model to generate synthetic chest X-ray images and uses these images to train a deep convolutional neural network (CNN) classifier based on the VGG-16 architecture. Then the model's performance was evaluated in classifying real chest X-ray images from the original dataset. The goal is to assess whether an auxiliary GAN can generate sufficiently realistic X-ray images such that a deep learning model trained on these data can accurately classify real X-ray images at scale.

Specifically, an auxiliary GAN was trained on a dataset of real patient chest X-rays to capture the underlying data, patterns and disease signatures. Then the trained model was used to generate a synthetic dataset of chest X-ray images modeled after the characteristics of the training set. The VGG-16 model was then trained and validated on GAN-generated synthetic images. Finally, the performance of the VGG-16 model was assessed on the original reserved patient dataset of real chest X-rays. By comparing the classification accuracy between the real and synthetic test sets, it was quantitatively evaluated how precisely the auxiliary GAN-replicated features are essential for accurate VGG-16 classification.

The study's methods and evaluation approach provide insights into the efficacy of auxiliary GANs in producing generalizable and useful synthetic medical training data. More broadly, the study contributes techniques to precisely analyze the generative model performance for clinical applications.

2. RELATED WORK
The use of generative models, such as generative adversarial networks (GANs), has rapidly grown to synthesize realistic medical images that can serve as training data for downstream applications. Prior studies have developed GANs for generating synthetic chest X-rays specifically to augment small real-world radiology datasets. For example, Madani et al. built a deep convolutional GAN able to generate 128x128 thoracic disease images shown to fool radiologists at a high rate.[1] Similarly, Guibas et al. developed a progressively growing GAN to output 256 × 256 chest X-rays with high fidelity. [2] These studies demonstrate chest X-ray synthesis as a valuable expansion of the limited medical imaging data.

A major challenge is the evaluation of the accuracy of GAN-generated medical images, which is critical for clinical applicability. Recent research has focused on improving the evaluation and quantification of the GAN performance for medical data. Salehinejad et al. evaluated chest X-ray GAN models by assessing generated images using a centroid distance metric against segmented lungs from real data.[3] However, few studies have examined downstream model performance using GAN-generated data as a proxy for precision and usefulness. This work addresses this gap by evaluating the classification fidelity of a common deep learning model, VGG-16, using synthesized X-ray data.

More broadly, prior literature has established deep CNNs like VGG-16 as highly effective for natural and medical image classification with proper tuning and training.[4] VGG-16 represents a strong deep CNN architecture validated for radiographic thoracic diagnosis.[5] Building on these works, this study employed VGG-16 as a reliable benchmark CNN for assessing chest X-ray data quality and features through our GAN evaluation approach. This study aimed to provide additional techniques for the precise analysis of GAN performance in healthcare contexts.

3. DATASET
In this research, the dataset played a pivotal role in training a Generative Adversarial Network (GAN) for conditional image generation focused on medical images, specifically chest X-rays for pneumonia detection. The dataset, sourced from the Kaggle Chest X-ray Pneumonia dataset [6], consists of two classes:

'NORMAL' for healthy individuals and 'PNEUMONIA' for those with pneumonia. This dataset was further divided into training (88 percent), testing (11 percent), and validation (1 percent) datasets. To enhance dataset preprocessing, a custom Python class named 'ReadDataset' was implemented. This class not only organizes the dataset path, labels, and desired image shape, but also provides methods to retrieve and read the images efficiently.

Preprocessing involves resizing the images to a specified dimension (64x64 pixels), converting them to the RGB color space, and normalizing pixel values between 0 and 1. In addition, the class employs Pathlib and OpenCV libraries to handle file paths and image reading. The resulting dataset, comprising 5216 images with corresponding labels, forms the foundation for training GAN. The utilization of Mean Squared Error (MSE) loss function in GAN architecture is crucial for addressing cognitive quality concerns and guiding the generator to focus on capturing distinctive features of both healthy and pneumonia-afflicted individuals. For testing, covid pneumonia normal chest Xray images, pneumonia Xray images, chest Xray pneumonia datasets sourced from Kaggle were utilized.

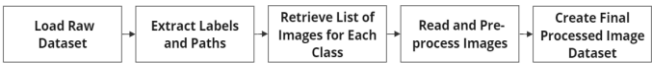

Figure 1: Pipeline for data pre processing

4. IMPLEMENTATION

To ensure the veracity of the generated medical images, a meticulous evaluation process was employed. This involves the creation of a substantial dataset comprising 30,000 samples enriched with random noise and diverse labels simulating various pathological conditions using the generator network. Subsequently, a neural network trained on this newly generated dataset was applied to assess the original dataset to accurately distinguish between healthy and pathological instances. This validation mechanism ensures that the generator not only produces visually accurate images but also captures nuanced pathological features aligned with real-world medical conditions. Furthermore, to achieve robust image classification, A pretrained VGG16 model was utilized as a feature extractor and subsequently fine-tuned for binary classification. The compiled model utilized binary cross-entropy loss and Adam optimizer, undergoing 60 epochs of training with a batch size of 64. By incorporating validation on a subset (20%) of the generated images and implementing an early stopping callback to prevent overfitting, these measures contribute to the reliability and diagnostic accuracy of the generated medical images in a formal and rigorous manner.

4.1 Generative Adversarial Network

Generative adversarial networks (GANs) are deep-learning architectures that are widely used in various machine-learning tasks, mainly for synthetic data generation. GANs consist of two neural networks that compete with each other using deep learning methods to artificially create data similar to the original dataset. GANs use a cooperative zero-sum game framework for learning, where one person's loss is the other person's gain. GANs consist of two components: a generator and a discriminator. The aim of the generator is to produce data that can be easily mistaken for the actual data. The discriminator uses the data produced by the generator as input. The goal of the discriminator is to distinguish between the data produced by the generator and actual data.

4.1.1 Generator.

In our pursuit of advancing Generative Adversarial Networks (GANs) for medical image synthesis, the crux lies in the intricacies of the generator architecture. This architectural structure serves as a force that translates random noise and essential diagnostic information into coherent visual representations.

The generator begins its orchestration with a concatenation of input noise and labeled data, creating a hybrid input that guides the creative process. A dense layer of 1024 units acts as the inception point, fostering a deep understanding of the information at hand. The subsequent dense layer, equipped with kernel regularization, unfolds into a three-dimensional space ($8 \times 8 \times 256$), setting the stage for detailed feature extraction.

A symphony of Conv2DTranspose layers follows, progressively upsampling the input to craft realistic and intricate medical images. Here, LayerNormalization is strategically incorporated to maintain stability and diversity in the generated images, a crucial step to avoid common pitfalls like mode collapse. ReLU activations infuse vitality into the model, allowing it to capture complex patterns and subtle nuances within the medical imagery.

The final crescendo involves a Conv2DTranspose layer generating a three-channel output (representing RGB) and an activation using the sigmoid function. This carefully designed architecture not only ensures the creation of visually authentic medical images but also contributes to a balanced and dynamic interplay between the generator and discriminator.

4.1.2 Discriminator.

In the pursuit of refining GANs for medical image synthesis, the focus was also on the discriminator architecture. This crucial component not only authenticates the generated images but also classifies them based on distinct health conditions.

The discriminator unfolds through the convolutional 2D layers by employing Leaky ReLU activation for feature extraction. Varying the filter sizes and strides captured both the local and global image patterns. Flattening precedes dense layers, outputting judgments on image authenticity and branching into an auxiliary classification pathway for health conditions. Regularization techniques, such as kernel regularization and dropout, enhance robustness, with Softmax activation providing meaningful probability distributions. This discriminator, which is pivotal in the Auxiliary GAN framework, excels in discernment and health classification. Its integration ensures the synthesis of contextually relevant medical images, marking a significant stride in the GAN capabilities for medical applications.

4.1.3 Auxiliary GAN.

In this research on Generative Adversarial Networks (GANs) for medical image synthesis, we adopt Auxiliary Classifier GAN (ACGAN) with carefully selected parameters. The ACGAN is configured with a learning rate (eta) of 0.0001, trained over 2000 epochs, and processed in batches of 16 samples. The generator, working with a latent space of 100, combines random noise and diagnostic information to create medical images. Simultaneously, the discriminator, with a convolutional kernel of size 5, evaluates the authenticity of generated images and provides auxiliary classifications related to specific health conditions. The interaction between these components, driven by the set parameters, reflects our commitment to developing a robust framework for generating coherent and clinically relevant medical images over an extended training period. The regularization technique, represented by a weight decay of 6e-9, contributes to the model's stability, emphasizing our approach to advancing GAN applications in medical imaging.

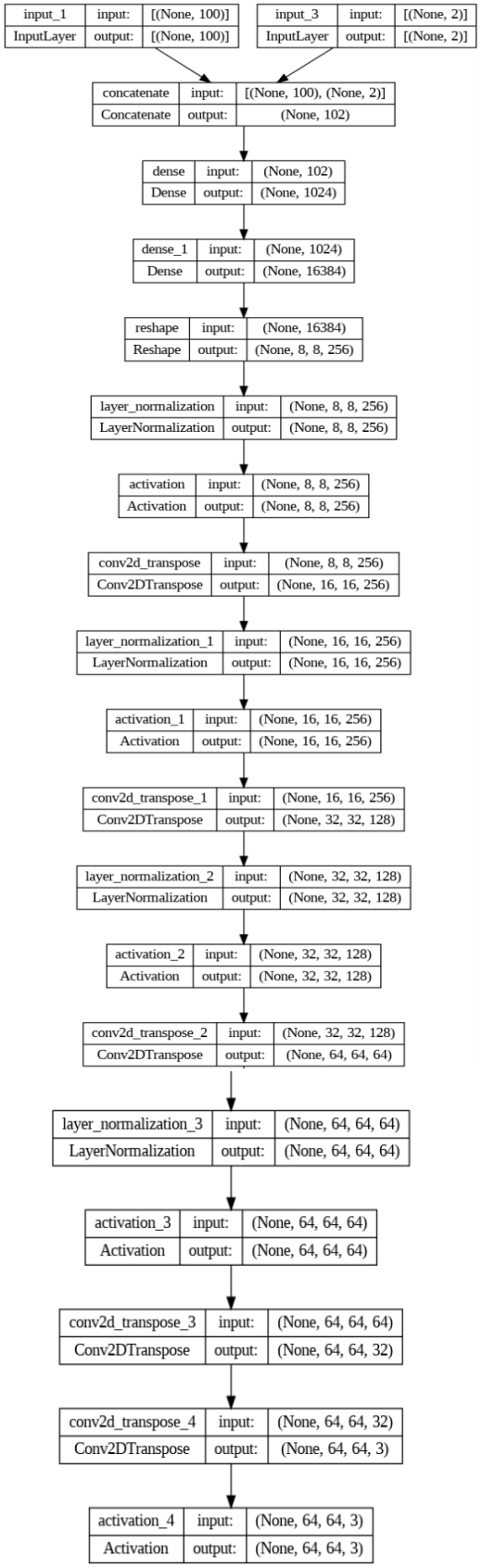

Figure 2: Architecture of the Generator

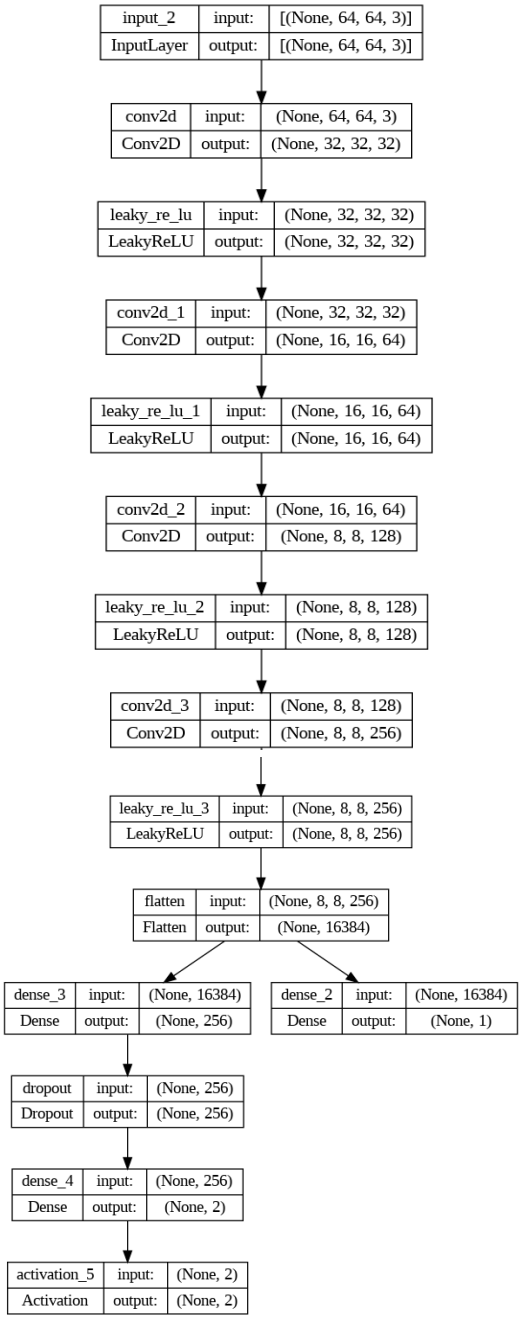

Figure 3: Architecture of the Discriminator

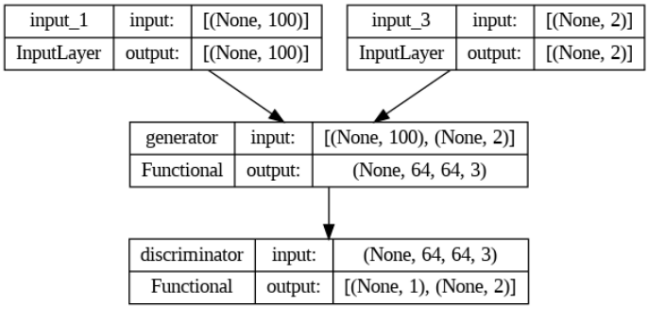

Figure 4: Auxiliary GAN architecture

Given our focus on a pathological condition, employing a larger kernel size proves instrumental in examining the interplay between core regions and their surroundings. This approach enhances the ability to detect gradients that signify the presence of either a pathological or healthy condition. Essentially, it aids in assessing whether a specific area, based on its location, can serve as a reliable criterion for identifying pneumonia. Additionally,

opting for LayerNormalization instead of BatchNormalization proves advantageous in preventing mode collapse during image generation by diversifying the generated images.

LayerNormalization's nuanced normalization at the filter level within the layer contributes to the stability of the generator, ensuring a richer and more varied set of generated images

4.2 EPOCH-WISE EVOLUTION OF GAN OUTPUTS

```
Epoch: 0
discriminator loss: [tag: 0.4942171573638916, labels: 0.7000121474266052], generator loss: [tag: 0.1631150245666504, labels: 0.66
3743793964386]
1/1 [==============================] - 0s 20ms/step
```

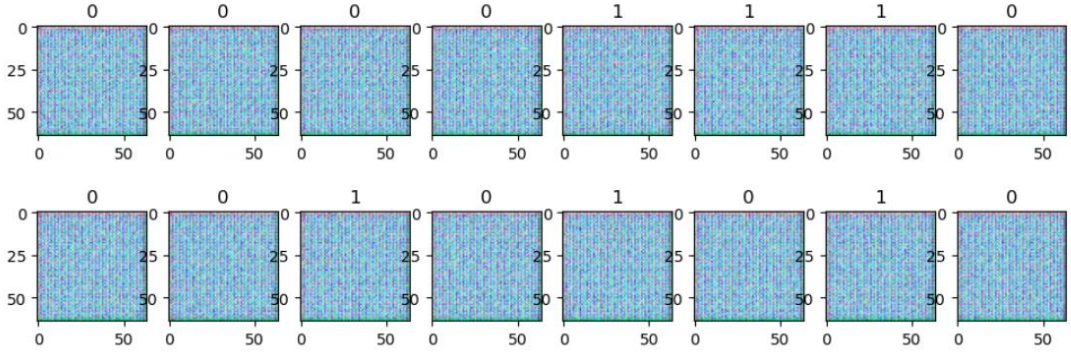

Figure 5: Outputs from epoch 0 of the GAN

```
Epoch: 1000
discriminator loss: [tag: 0.27188217639923096, labels: 0.021886713802814484], generator loss: [tag: 0.18630945682525635, labels:
0.00786643661558628]
1/1 [==============================] - 0s 19ms/step
```

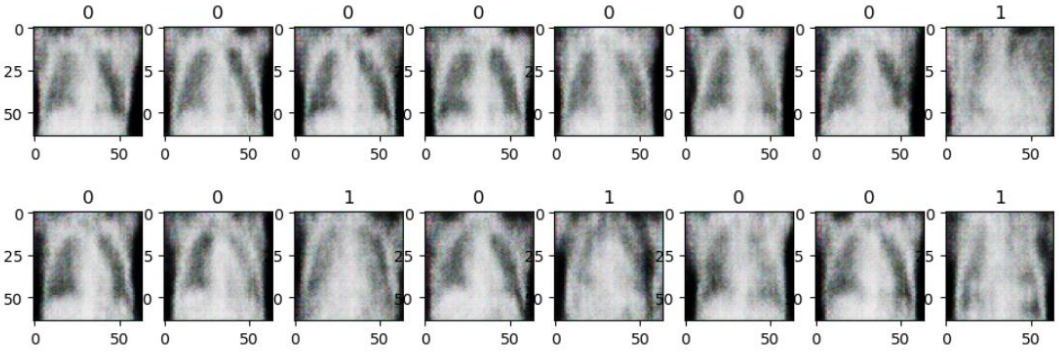

Figure 6: Outputs from epoch 1000 of the GAN

```
Epoch: 1900
discriminator loss: [tag: 0.22534728050231934, labels: 0.016925722360610962], generator loss: [tag: 0.35416722297668457, labels:
0.0029466827400028706]
1/1 [==============================] - 0s 20ms/step
```

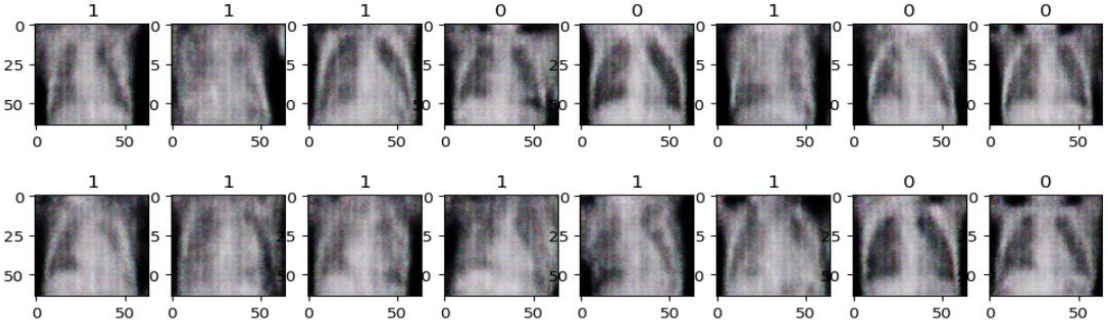

Figure 7: Outputs from epoch 1900 of the GAN

4.3 CLASSIFIER

To ensure the fidelity of the generated images, a meticulous evaluation process is implemented. A substantial dataset, comprising 30,000 samples, is generated using the generator network. This dataset is enriched with random noise and corresponding labels to simulate diverse pathological conditions. Leveraging this newly generated dataset, a neural network is employed to classify the images, effectively discerning between healthy and pathological instances. Subsequently, the trained classification neural network is applied to evaluate the fundamental images present in the original dataset. This evaluative step is crucial as it gauges the effectiveness of the learned characteristics from the generated images in accurately classifying the inherent pathology within the baseline dataset. In essence, this approach serves as a validation mechanism, ensuring that the generator not only produces visually accurate images but also captures the nuanced pathological features that align with real-world medical conditions.

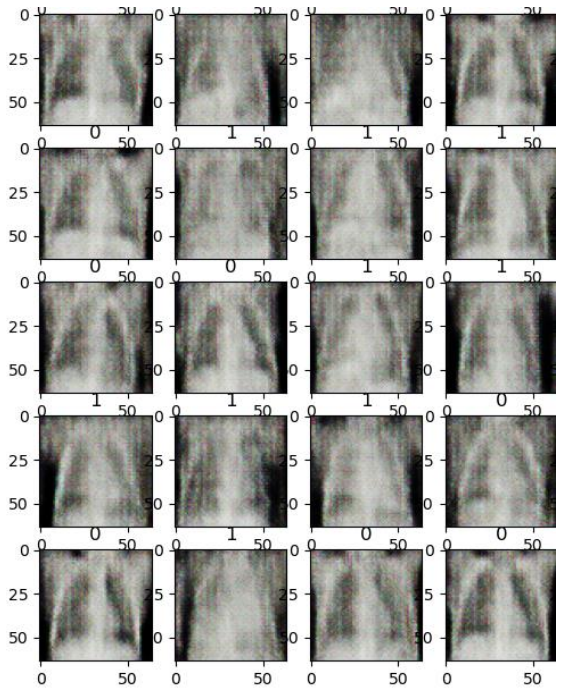

Figure 8: Synthetic data from the generator

In pursuit of robust image classification for the generated dataset, a pre-trained VGG16 model is leveraged as a feature extractor. The base VGG16 model, initialized without pre-trained weights, processes input images of size (64, 64, 3) and employs max-pooling for spatial down-sampling while excluding the top classification layer. Subsequent to the feature extraction, additional layers are appended to fine-tune the model for binary classification. A dropout layer with a rate of 0.4 is introduced to decrease overfitting, followed by densely connected layers for feature refinement. Batch normalization ensures a stable and accelerated convergence during training, while Leaky ReLU activations with a modest negative slope of 0.2 introduce non-linearity to the model. The final dense layer, activated by the sigmoid function, yields binary classification results.

The model is compiled using binary cross-entropy loss and Adam optimizer with a learning rate of 0.00001. During training, the model is validated on a subset (20%) of the generated images to assess its generalization capabilities. The training process spans 60 epochs with a batch size of 64, and an early-stopping callback halts training if the validation loss fails to improve for two consecutive epochs, thereby preventing overfitting. This comprehensive classification model, built on top of the VGG16 feature extractor, is instrumental in evaluating the quality of generated images and validating the learned characteristics against the original dataset.

5. RESULTS

The plotted metrics depict the training dynamics and generalization of our neural network on images generated by GAN. The descending training loss reflects parameter optimization on the training set, while the validation loss gauges the model's ability to generalize to unseen data. Minimal divergence from the training loss curve indicates a robust generalization. These metrics offer insights into model convergence, potential overfitting, and the network's effectiveness in capturing underlying patterns within the generated medical images.

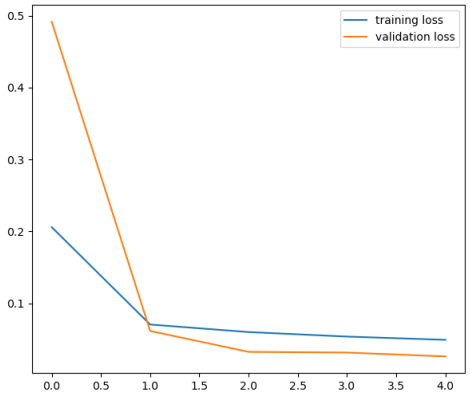

Figure 9: The results obtained from the classifier

Following training on images generated by the generator, the neural network undergoes rigorous testing on the fundamental images within the dataset. Multiple evaluation metrics are employed to assess the generative adversarial network's efficacy in capturing intrinsic features characteristic of each class. This scrutiny also extends to the secondary classification network, examining its capability to extract features present in the generated images. The pivotal question revolves around the transferability of attributes extracted from generated images to the original dataset. This assessment serves as a pivotal step in scrutinizing the authenticity of the generated images and verifying whether the focus indeed encapsulates the specific cases indicative of pneumonia or its absence, as discerned from the X-ray images.

Table 1. The results of the evaluation metrics

	precision	**recall**	**f1-score**	**support**
0	0.99	0.64	0.78	2069
1	0.81	1.00	0.89	3147
accuracy			0.85	5216
macro avg	0.90	0.82	0.83	5216
weighted avg	0.88	0.85	0.85	5216

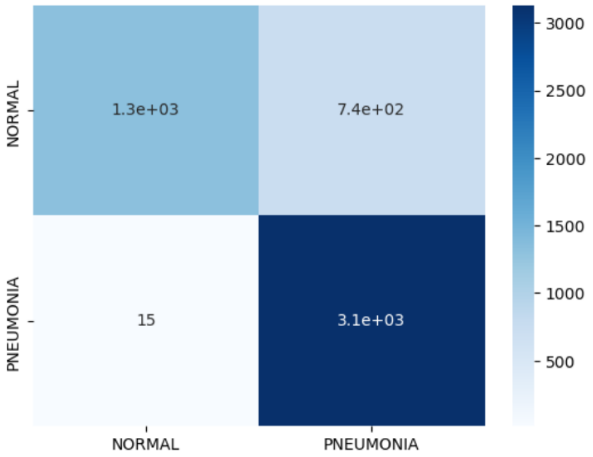

Figure 10: A confusion matrix of classifier's output

Table 2. Comparison of results across different datasets

Dataset	GAN accuracy	GAN + RL Accuracy
Paul Mooney's Chest X-Ray Images (Pneumonia)	74.2906	89.8773
Covid-pneumonia-normal-chest-xray-images	76.0966	87.0905
Pneumonia-xray-images	84.5900	89.4306

6. FUTURE WORKS

In paving the way for future advancements, the seamless integration of the classifier into a reinforcement learning framework becomes a paramount consideration. A key contribution lies in the formulation of the 'model fn' function, strategically designed to process observations using the classifier model. Within the domain of Deep Q Networks (DQN), an integral component, the Q-network, is meticulously crafted with specific layer configurations. Emphasizing efficiency in training, Adam optimizer, wielding a learning rate of 0.001, is employed to propel the learning dynamics.

Looking ahead, the initialization and training loop of the agent mark essential facets of the research framework. The DQN agent is primed with a robust architecture, comprising the Q-network, Adam optimizer, and a time step counter to monitor learning progress. The training trajectory unfolds across a predefined number of iterations, currently set at 100 for illustrative purposes. After each iteration, the environment undergoes a reset, prompting the agent to engage in action selection based on its evolving policy. These trajectories are systematically cataloged in the agent's replay buffer, laying the groundwork for subsequent training endeavors.

7. CONCLUSION

This study explores the application of Generative Adversarial Networks (GANs) in synthesizing realistic medical images, with a focus on chest X-rays. Utilizing an auxiliary GAN, synthetic chest X-rays are generated and employed to train a VGG-16 Convolutional Neural Network (CNN) classifier. The main objective is to evaluate the ability of auxiliary GAN to produce realistic images, allowing the trained CNN to accurately classify real X-ray images at scale. The methodology involves training auxiliary GAN on real patient chest X-rays, generating a synthetic dataset that mimics the original training set. The subsequent training and validation of the VGG-16 model on these GAN-generated images enable a quantitative analysis of classification accuracy, providing insights into GAN's efficacy in replicating features essential for accurate CNN performance.

In addition to classifier results, the study includes a detailed analysis of training dynamics and generalization of the neural network on GAN-generated images. Plotted metrics, such as training and validation loss, offer indicators of model convergence, potential overfitting, and the network's effectiveness in capturing underlying patterns within the synthetic medical images. This comprehensive approach contributes insights into the precise analysis of generative model performance, addressing the challenge of quantitative evaluation in the context of synthetic medical training data, and providing valuable implications for clinician applications in medical image synthesis and diagnostics.

8. REFERENCES

[1] A. Madani, M. Moradi, A. Karargyris, and T. Syeda-Mahmood, "Chest x-ray generation and data augmentation for cardiovascular abnormality classification," in 2018 IEEE 31st International Symposium on Computer-Based Medical Systems (CBMS), 2018, pp. 360–364.

[2] L. Guibas, T. Virdi, and P. Li, "Synthetic Medical Images from Dual Generative Adversarial Networks," arXiv:1709.01872 [cs], Sep. 2017, Accessed: Dec. 05, 2023. [Online]. Available: http://arxiv.org/abs/1709.01872

[3] H. Salehinejad, S. Valaee, T. Dowdell, E. Colak, and J. Barfett, "Generalization of Deep Neural Networks for Chest Pathology Classification in X-Rays Using Generative Adversarial Networks," in 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2018, pp. 990–994.

[4] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," in Advances in Neural Information Processing Systems, vol. 25, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2012, pp. 1097–1105.

[5] X. Wang et al., "ChestX-ray8: Hospital-scale Chest X-ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases," in 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 3462–3471.

[6] Paul Mooney, (2018, January). Chest X-Ray Images (Pneumonia), Version 2. Retrieved December 1, 2023 from https://www.kaggle.com/datasets/paultimothymooney/chest-xray-pneumonia