LEHIGH UNIVERSITY
M.S. IN DATA SCIENCE

DSCI 441: Statistical and Machine Learning

Assignment Report

Substantiating diagnostic capabilities using Anomaly Detection with Generative Adversarial Networks

---

Advisor: Prof. Yari Masoud

BETHLEHEM, PENNSYLVANIA, MAY 2023

Contents

1 Motivation and Background

With increased resource utilization due to increased burden on patients as well as healthcare system, detection of diagnosis using automated machine learning techniques is the need of the hour. We have come a long way in supporting physicians and health policy makers by identifying brand-new biomarkers for myriad of diseases. As data scientists, we still have a longer road ahead in matching physician expertise for detecting different diagnoses using algorithms. Using deep learning and medical images like X-Ray, in this study we attempt to devise a methodology to identify tuberculosis in patients. We always find more data records with absence of disease than presence and this statistically pose a major challenge to develop a supervised disease detection model. Therefore, in this study we turn to unsupervised techniques to develop an anomaly detection model using Generative Networks. We believe that with enough samples of Chest X-Ray images for healthy patients, we can train a Generative Adversarial Network (GAN) to detect the disease. The disease instance should lie in the outlier zone of the trained GAN model as it could only regenerates image of healthy patients with minimal reconstruction loss. Eventually, we build a "Streamlit" Dashboard which uses a new Chest X-Ray image to predict if the patient has Tuberculosis. This product could assist physician in making more informed and faster decision. A similar product could even be provided to patients allowing them to test presence or possibility of having a certain diagnosis in real time, allowing patients to become aware and eventually supporting the common cause of early diagnosis and early medical actions.

2 Recent work done in this domain

Anomalies arise due to various reasons such as data errors or data noises but sometimes indicate a new process that was previously unseen. Thus, anomaly detection is a crucial task, especially in medical image processing. These anomalies can be used to identify unhealthy patients as unhealthy patients can be considered as an outlier in complete set healthy and unhealthy patients. Many researchers tended to employ deep learning to detect abnormalities in images, due to the proliferation of deep neural networks, with unprecedented results across various applications. It can also deal with complicated features such as regions of interest points by examining every pixel in an image. In fact, deep learning-based anomaly detection have gained prominence and have been applied to various tasks, with the help of the technologies increasingly popular in the medical sector [7] [13]. This is because deep learning overcomes the issue of data being imbalanced, which may result in a bias towards the majority group (i.e., the negative case). Since the medical images for the negative cases are more than the positive ones, we believe that anomaly detection can be considered a better technique to be adopted than the binary classification Machine learning has revolutionized health science research, especially in neuro, respiratory and cardiovascular area [10], such as by modeling interactions between whole brain genomics/imaging [8] and identifying Alzheimer's disease (AD)-related proteins or using Chest X-Ray images to identify Covid-19 [11]. Especially, deep learning can achieve accurate computer-assisted diagnosis when large-scale annotated training samples are available. In medical imaging, unfortunately, preparing such massive annotated datasets is often unfeasible; to tackle this pervasive problem, researchers have proposed various data augmentation techniques, including generative adversarial network (GAN)-based ones [4], [5]. However, even exploiting these techniques, supervised learning still requires many images with pathological features, even for rare diseases, to make a reliable diagnosis; nevertheless, it can only detect already-learned specific pathologies. In this regard, as physicians notice previously unseen anomaly examples using prior information on healthy body structure, unsupervised anomaly detection methods leveraging only large-scale healthy images can discover and alert overlooked diseases when their generalization fails.

Towards this, researchers reconstructed a single medical image via GANs [9], autoencoders (AEs) , or com-

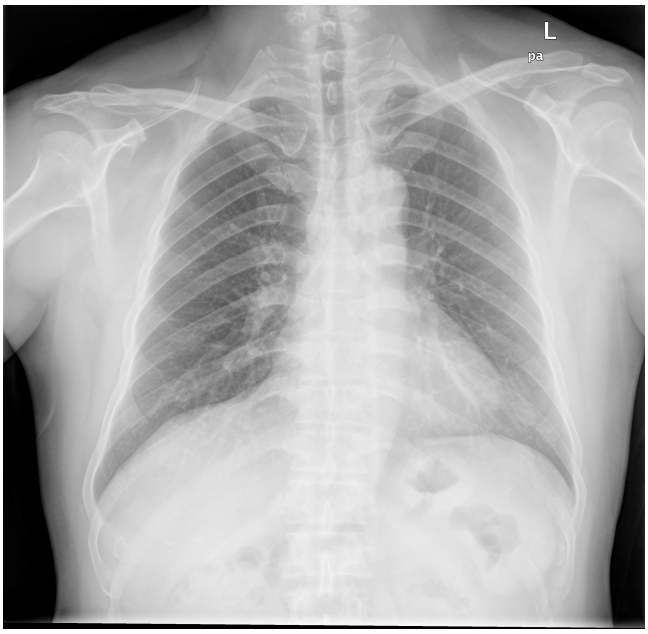

Figure 1: A random chest X-Ray image from the Shenzhen dataset of size $3000X2919$

bining them, since GANs can generate realistic images and AEs, especially variational AEs (VAEs), can directly map data onto its latent representation [3]; then, unseen images were scored by comparing them with reconstructed ones to discriminate a pathological image distribution (i.e., outliers either in the learned feature space or from high reconstruction loss).

3 Datasets Used

To train an exhaustive AnoGAN, we made an attempt to train the ANoGAN on large set of Chest X-Ray images coming from different sources. In this study we primarily used three publicly available data sources of chest X-Ray images with labels provided for each image in the data sets.

3.1 China Set - The Shenzhen set

The standard digital image database for Tuberculosis is created by the National Library of Medicine, Maryland, USA in collaboration with Shenzhen No.3 People's Hospital, Guangdong Medical College, Shenzhen, China [6]. The Chest X-rays are from out-patient clinics, and were captured as part of the daily routine using Philips DR Digital Diagnose systems. The dataset was de-identified by the data providers and was exempted from IRB review at their institutions, where label '0' represents the normal and '1' represents the abnormal lung. See Figure 1.

3.2 Montgomery County - Chest X ray

The standard digital image database for Tuberculosis is created by the National Library of Medicine in collaboration with the Department of Health and Human Services, Montgomery County, Maryland, USA [6]. The set contains data from X-rays collected under Montgomery County's Tuberculosis screening program. Image file names are coded as $MCUCXRXX0/1.png$, where '0' represents the normal and '1' represents the abnormal lung.

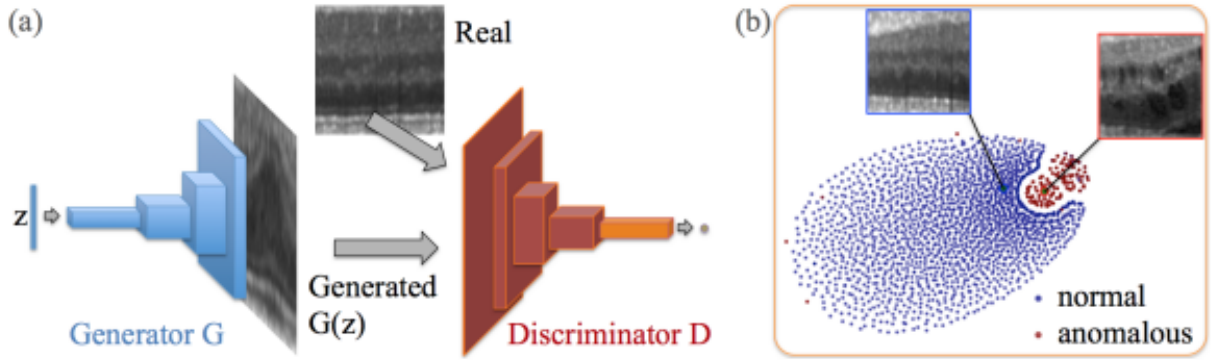

Figure 2: (a) Deep convolutional generative adversarial network. (b) t-SNE embedding of normal (blue) and anomalous (red) images on the feature representation of the last convolution layer (orange in (a)) of the discriminator. [14]

3.3 TBX11K Set

In the paper [7], we contribute to the community with a large-scale Tuberculosis X-ray (TBX11K) dataset, through the long-term cooperation with major hospitals. This new dataset is superior to previous CTD datasets in the following aspects: i) Unlike previous datasets [6, 18] that only contain several tens/hundreds of X-ray images, TBX11K has 11,200 images that are about 17× larger than the existing largest dataset, i.e., Shenzhen dataset [18], so that TBX11K makes it possible to train very deep CNNs; Each X-ray image in TBX11K is tested using the golden standard (i.e., diagnostic microbiology) and then annotated by experienced radiologists from major hospitals. TBX11K dataset has been de-identified by the data providers and exempted by relevant institutions, so it can be made publicly available to promote future CTD research.

4 AnoGAN for Anomaly Detection

To identify anomalies, we learn a model representing normal anatomical variability based on GANs [14]. This method trains a generative model, and a discriminator to distinguish between generated and real data simultaneously 2. Instead of a single cost function optimization, it aims at the Nash equilibrium of costs, increasing the representative power and specificity of the generative model, while at the same time becoming more accurate in classifying real- from generated data and improving the corresponding feature mapping. A GAN consists of two adversarial modules, a generator G and a discriminator D. The generator G learns a distribution p_g over data x via a mapping G(z) of samples z, 1D vectors of uniformly distributed input noise sampled from latent space Z, to 2D images in the image space manifold X , which is populated by healthy examples. The discriminator output D(·) can be interpreted as probability that the given input to the discriminator D was a real image x sampled from training data X or generated G(z) by the generator G. D and G are simultaneously optimized through the following two-player minimax game with value function V (G, D).

$$\min_G \max_D V(D,G) = E_{x\sim p_{data}(x)}[logD(x)] + E_{z\sim p_z}[log(1\text{-}D(G(z)))] \quad (1)$$

The discriminator is trained to maximize the probability of assigning real training examples the *real* and samples from pg the $fake$ label. The generator G is simultaneously trained to fool D via minimizing $V(G) =$

$log(1D(G(z)))$, which is equivalent to maximizing

$$V(G) = D(G(z)) \tag{2}$$

During adversarial training the generator improves in generating realistic images and the discriminator progresses in correctly identifying real and generated images.

4.1 Mapping new Images to the Latent Space

Both components of the trained GAN, the discriminator D and the generator G, are utilized to adapt the coefficients of z via backpropagation. In the following, we give a detailed description of both components of the loss function. Residual Loss The residual loss measures the visual dissimilarity between query image x and generated image G(z) in the image space and is defined by LR(z) = X|x G(z)|. (3) Under the assumption of a perfect generator G and a perfect mapping to latent space, for an ideal normal query case, images x and G(z) are identical. In this case, the residual loss is zero.

$$L_R(z_\gamma) = \sum | \, x - G(z_\gamma) \, | \tag{3}$$

Discrimination Loss For image inpainting, Yeh et al. [15] based the computation of the discrimination loss $L_{\hat{D}}(z_\gamma)$ on the discriminator output by feeding the generated image G(z) into the discriminator $L_{\hat{D}}(z_\gamma) = \sigma(D(G(z)), \alpha)$, where is the sigmoid cross entropy, which defined the discriminator loss of real images during adversarial training, with logits $D(G(z_\gamma))$ and targets = 1

$$L_D(z_\gamma) = \sum | \, f(x) - f(G(z_\gamma)) \, | \tag{4}$$

4.1.1 Detection of Anomalies

During anomaly identification in new data we evaluate the new query image x as being a normal or anomalous image. Our loss function (Eq. (5)), used for mapping to the latent space, evaluates in every update iteration the compatibility of generated images $G(z_\gamma)$ with images, seen during adversarial training. Thus, an anomaly score, which expresses the fit of a query image x to the model of normal images, can be directly derived from the mapping loss function (Eq. (5)): A(x) = (1) R(x) + D(x), (6) where the residual score R(x) and the discrimination score D(x) are defined by the residual loss $L_R(z_\gamma)$ and the discrimination loss $L_D(z_\gamma)$ at the last update iteration of the mapping procedure to the latent space, respectively.

$$L(z_\gamma) = \sum | \, (1\text{-}\lambda)L_R(z_\gamma) - \lambda L_D(z_\gamma) \, | \tag{5}$$

5 AnoGAN for Tuberculosis detection

AnoGAN in the original paper [9] was implemented to detect anomalies in clinical high resolution SD-OCT volumes of the retina with 49 B-scans (representing an image slice in zx-plane) per volume and total volume resolutions of 496×512×49 voxels in z-, x-, and y direction, respectively. In this exercise we intend to extend the application of using DCGAN as anomaly detection model for Tuberculosis. This study is based on understanding that just like any medical imaging AnoGAN can be used to identify abnormalities among Tuberculosis patients. The abnormalities in Chest X-Ray images can be theoretically used to identify active Tb. Following abnormalities are evident in chest X-Ray images -

Figure 3: Chest x-ray of Ghon's complex of active tuberculosis

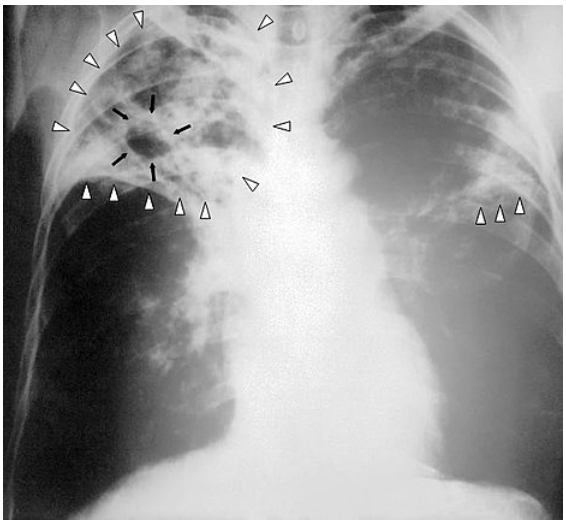

Figure 4: Chest X-ray of a person with advanced tuberculosis: Infection in both lungs is marked by white arrowheads, and the formation of a cavity is marked by black arrows.

Figure 5: Chest x-ray showing dense opacity pleural effusion in the lower left lung of primary pulmonary TB.

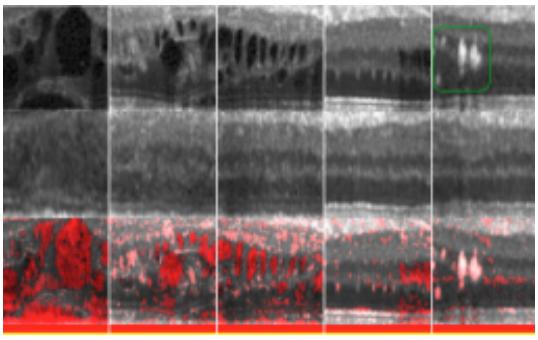

Figure 6: Abnormalities detection in the original AnoGAN paper [9]

1. Infiltrate or consolidation - Opacification of airspaces within the lung parenchyma. Consolidation or infiltrate can be dense or patchy and might have irregular, ill-defined, or hazy borders, see figure 3.
2. Any cavitary lesion - Lucency (darkened area) within the lung parenchyma, with or without irregular margins that might be surrounded by an area of airspace consolidation or infiltrates, or by nodular or fibrotic (reticular) densities, or both. The walls surrounding the lucent area can be thick or thin. Calcification can exist around a cavity see figure 4.
3. Nodule with poorly defined margins - Round density within the lung parenchyma, also called a tuberculoma. Nodules included in this category are those with margins that are indistinct or poorly defined (tree-in-bud sign). The surrounding haziness can be either subtle or readily apparent and suggests coexisting airspace consolidation.
4. Pleural effusion - Presence of a significant amount of fluid within the pleural space. This finding must be distinguished from blunting of the costophrenic angle, which may or may not represent a small amount of fluid within the pleural space (except in children when even minor blunting must be considered a finding that can suggest active TB) see figure 5.
5. Hilar or mediastinal lymphadenopathy (bihilar lymphadenopathy) - Enlargement of lymph nodes in one or both hila or within the mediastinum, with or without associated atelectasis or consolidation.
6. Linear, interstitial disease (in children only) - Prominence of linear, interstitial (septal) markings.
7. Other - Any other finding suggestive of active TB, such as miliary TB. Miliary findings are nodules of millet size (1 to 2 millimeters) distributed throughout the parenchyma.

These abnormalities are very similar to what was detected in the original AnoGAN paper 6 and hence we may be able to accurately detect and locate pulmonary tuberculosis using AnoGAN. There are multiple supervised deep learning approaches which attempts to classify Tuberculosis patients [1]. This will be the first unsupervised deep learning GAN based approach which attempts to classify Tuberculosis patients.

6 Data Analysis Approach

As discussed, the AnoGAN trains a generator (G) to take an input noise vector (z) and produce an image which follows the distribution learnt from healthy images and when such generated image is compared with a real chest X-Ray of a patient with Tuberculosis, a higher loss score is generated in comparison to the loss score generated when the generated image is compared with chest X-Ray of a healthy patient. The normal/healthy

Training Data			
Patients	TBX11K	Shenzhen	Montgomery County
Sick or Tb	0	0	0
Healthy	3,377	326	0
Validation Data (To decide threshold)			
Tb	101	201	0
Testing Data			
Healthy	348	75	80
Tb	699	135	58
Sick and Non-Tb	3,800	0	0

Table 1: Datasets used for training and validation

Hyper-parameters	Explored	Used
Length of Noise vector (Z)	[10, 20, 50]	10
Image dimensions	28X28; 64X64; 256X256	256X256
Proportion of discriminator and residual loss	$\lambda = [0.5, 0.9, 1]$	0.9
Optimizer	Adams, RMSprop	RMSprop
Learning rate	$D = [0.0001, 0.004], G = [0.0001, 0.002]$	$D = 0.0004, G = 0.0002$

Table 2: Hyper-parameter tuning for AnoGAN model

images from all the TBX11K and Shenzhen data sets ($N = 3,377$) were split were used to train the model and validation set were Tuberculosis patients from TBX11K and Shenzhen Datasets ($N = 326$). See Table 1 to understand split of train and validation set across multiple data sets. Please note, during the time of training, no Chest X-Ray images of sick patients were used and at the time of validation healthy patients were not included to decide the thresholds.

6.1 Datasets used for training and testing

After the model was trained using the training healthy images and validation Tuberculosis images, the performance of the model was evaluated on healthy, tuberculosis and sick (Non-Tb) patients separately while using highlighted η as the threshold. See heading Testing Data of table, to understand

6.2 Image preprocessing

The Chest X-Ray images from different sources are bound to have different resolution (u,v) and for the model to be able to successfully make a prediction for all chest X-Ray images, the training X-Ray images were converted from (u, v) to (256, 256) and the GAN was trained on the fixed dimension of size (256, 256). All images were also converted from RGB to grayscale and hence input dimension of training X-Ray images was ($n_{samples}$, 256, 256, 1).

6.3 AnoGAN Hyper parameter tuning

As discussed in section 6, it is an unupervised approach of detecting anomalies which means that model is not supplied with labels for the input images. The Generator and discriminator are trained to achive a nash equilibrium and this process the Generator gets better at generating image patterns which can fool the discriminator. See Figure 7 to understand the functionality of AnoGAN. The left side of the diagram represent

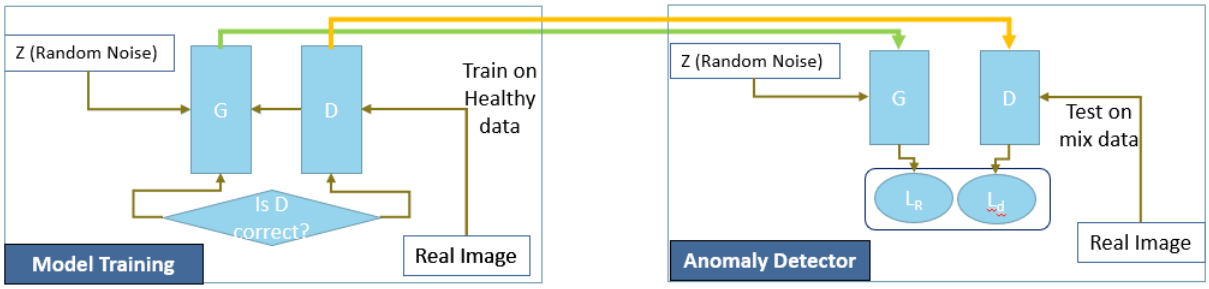

Figure 7: Diagram explaining the functionality of AnoGAN

the model training and the right side represents calculation of anomaly score using residual loss and discriminator loss. Either of these losses can be used as anomaly score. In this exercise, we used eq. 5 to calculate the final anomaly score with $\lambda = 0.1$ which translates the eq. to following. Similarly, other hyper-paramters like dimension of latent space (z_{dim}), size of the image and, learning rate of the model were tuned using grid search , see table 2.

$$A(x) = \sum | (0.9)R(x) - (0.1)D(x) | \tag{6}$$

6.4 Deciding threshold for anomaly detection

The model was iteratively retrained to achieve a minimum overlap between anomaly score distribution of healthy trained images and anomaly score distribution of 100 unhealthy test images. In simple words, once the model is trained, anomaly score for 3,377 healthy images and 302 unhealthy images were recorded and the distributions of both cases were observed to identify a threshold that separates the two distribution with highest accuracy. This distribution plot was observed by changing different hyperparameters in the model to achieve a model that best separates the anomaly scores of healthy and unhealthy set. See Figure 8 to get a sense of model performance, where "red" represents the anomaly score density plot for healthy patients and "blue" represents the anomaly score density plot for Tuberculosis patients, based on this plot the intersection of "red" and "blue" curve was used as the threshold values $\eta = 9.5X10^6$

6.5 Post processing and identification of Anomalous images

Test images were first resized from size (u, v) to (256, 256) to be used as an input to the trained GAN model. And, for each image the residual loss (L_R) and discriminant loss (L_D) are calculated and their weighted sum with $\lambda = 0.1$ is used as final anomaly. If the calculated anomaly score for a given image is greater than the decide threshold, the image is considered to be anomalous and the patient is probable to have Tuberculosis and if the anomaly score is less than the decided threshold, the patient is considered to be healthy.

7 Results and Discussion

After the GAN, was trained the model was used to calculate the anomaly score and as mentioned in (Section 6.4) $\eta = 9.5X10^6$ was used as the final threshold of the model to classify between Anomalous and Non-Anomalous. See Table 3 to view the performance of the model across the testing data of all three databases. Model performed best on Shenzhen dataset with 84.7% accuracy and F1-score of 85.7% and worst on Montgomery county dataset with accuracy of 55.1% and F1-score of 41.2%. As mentioned in previous sections,

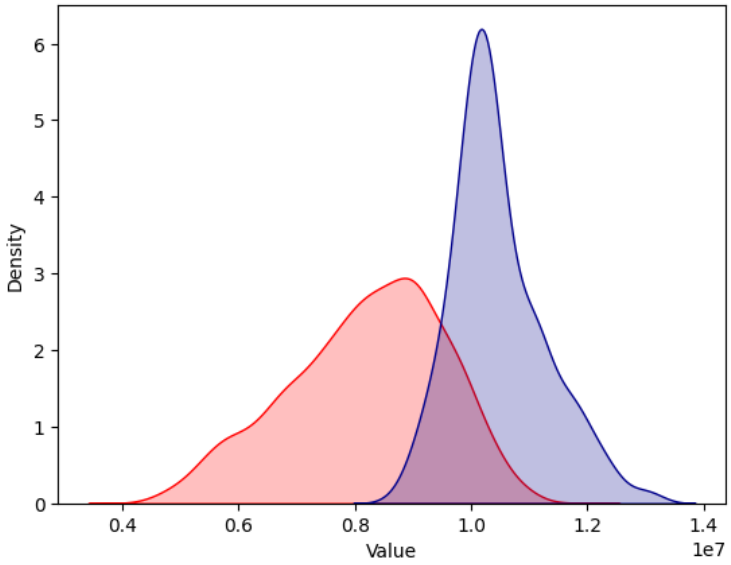

Figure 8: Validation plot to decide the threshold for Anomaly score

montgomery county dataset was not included among samples which were used to train the GAN model and hence highlighting the dependence of GAN on variance of the training samples. Like any machine learning model, the performance of GAN is often limited by the variety of samples included while training. Though the chest X-Ray images are almost symmetric in nature, data augmentation by adding images with different level of zoom may improve the model performance and augmentation using horizontal or vertical flips is expected to not effect the current model performance. Also, DCGAN when used for anomaly detection requires more computational power than most machine learning models when used on image datasets because of multiple fully connected layers involved in both generator and discriminator architecture. There was another challenge of stagnating losses while training which was handled by hyper-parameter tuning of the model as mentioned in section 6.3 and by saving model weights at regular checkpoints. When training was complete, the checkpoint with most appropriate generator and discriminator loss was chosen to validate the model performance on validation set. Thus most appropriate weights were chosen and same weights were used to make reliable prediction in the streamlit app [2]. Therefore, to achieve better performance with the given architecture, a larger dataset of healthy patients originating from different sources would be helpful and thus the GAN would require more number of GPU units to achieve steady learning.

7.1 GAN model capabilities

It is essential to check if the GAN model trained here is actually learning anatomical features which makes sense and the performance seen by the model across multiple datasets is not by chance and hence we evaluate the feature representation of the discriminator trained. Figure 9 highlights t-SNE embedding of features extracted by discriminator in the last layer of the discriminator architecture. It can be seen that there is almost a clear separation of t-SNE embedding of anomalous (un-healthy/ Tb) images and normal (healthy) images. Therefore, we can conclude that the discriminator is learning meaningful features that separates the two classes and model performance is not by chance. However, it is to be noted that GAN is still subject to training samples and poor generalization for chest X-Ray images obtained from different data sources.

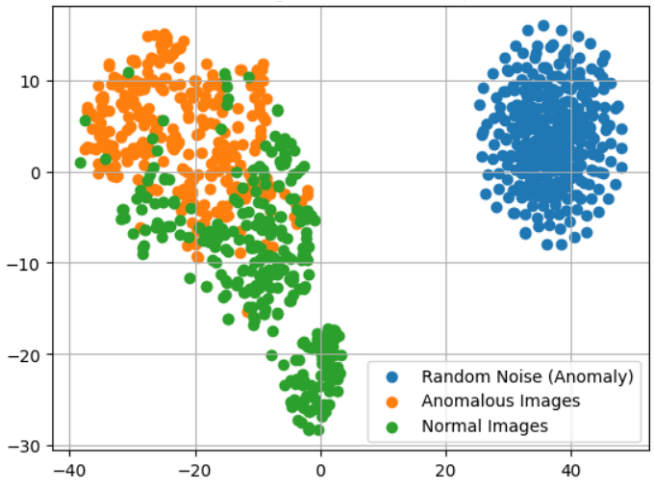

Figure 9: t-SNE embedding of normal (green) and anomalous (orange) images on the feature representation of the last convolution layer of the discriminator.

Testing Data		
Datasets	Accuracy	F1-Score
TBX11K	71.5%	65.8%
Shenzhen	84.7%	85.7%
Montgomery County	55.1%	41.2%

Table 3: Model Performance on Testing Datasets

References

[1] Evaluating effect of lung segmentation on disease prediction. https://medium.com/@abs422/evaluating-effect-of-lung-segmentation-using-u-net-on-disease-prediction-d208a767a829. Accessed: 2022-10-20.

[2] Tuberculosis prediction app - a gan based approach. https://abs422-dsci-441-gan-tb-streamlit-app-kp3tpv.streamlit.app/. Accessed: 2022-10-20.

[3] Xiaoran Chen and Ender Konukoglu. Unsupervised detection of lesions in brain MRI using constrained adversarial auto-encoders. *CoRR*, abs/1806.04972, 2018.

[4] Veronika Cheplygina, Marleen de Bruijne, and Josien PW Pluim. Not-so-supervised: a survey of semi-supervised, multi-instance, and transfer learning in medical image analysis. *Medical image analysis*, 54:280–296, 2019.

[5] Maayan Frid-Adar, Idit Diamant, Eyal Klang, Michal Amitai, Jacob Goldberger, and Hayit Greenspan. Gan-based synthetic medical image augmentation for increased cnn performance in liver lesion classification. *Neurocomputing*, 321:321–331, 2018.

[6] Stefan Jaeger, Sema Candemir, Sameer Antani, Yì-Xiáng J Wáng, Pu-Xuan Lu, and George Thoma. Two public chest x-ray datasets for computer-aided screening of pulmonary diseases. *Quantitative Imaging in Medicine and Surgery*, 4(6):475, 2014.

[7] Yun Liu, Yu-Huan Wu, Yunfeng Ban, Huifang Wang, and Ming-Ming Cheng. Rethinking computer-aided tuberculosis diagnosis. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2643–2652, 2020.

[8] Sarah E Medland, Neda Jahanshad, Benjamin M Neale, and Paul M Thompson. Whole-genome analyses of whole-brain data: working within an expanded search space. *Nature neuroscience*, 17(6):791–800, 2014.

[9] Thomas Schlegl, Philipp Seeböck, Sebastian M. Waldstein, Ursula Schmidt-Erfurth, and Georg Langs. Unsupervised anomaly detection with generative adversarial networks to guide marker discovery. *CoRR*, abs/1703.05921, 2017.

[10] Angela Serra, Paola Galdi, and Roberto Tagliaferri. Machine learning for bioinformatics and neuroimaging. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 8(5):e1248, 2018.

[11] Shivani Sharma and Shamik Tiwari. Covid-19 diagnosis using x-ray images and deep learning. In *2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS)*, pages 344–349, 2021.

[12] Wen Shi, Guohui Yan, Yamin Li, Haotian Li, Tingting Liu, Cong Sun, Guangbin Wang, Yi Zhang, Yu Zou, and Dan Wu. Fetal brain age estimation and anomaly detection using attention-based deep ensembles with uncertainty. *NeuroImage*, 223:117316, 2020.

[13] Shuaijing Xu, Hao Wu, and Rongfang Bie. Cxnet-m1: Anomaly detection on chest x-rays with image-based deep learning. *IEEE Access*, 7:4466–4477, 2019.

[14] Raymond A. Yeh, Chen Chen, Teck-Yian Lim, Mark Hasegawa-Johnson, and Minh N. Do. Semantic image inpainting with perceptual and contextual losses. *CoRR*, abs/1607.07539, 2016.

[15] Raymond A. Yeh, Chen Chen, Teck-Yian Lim, Mark Hasegawa-Johnson, and Minh N. Do. Semantic image inpainting with perceptual and contextual losses. *CoRR*, abs/1607.07539, 2016.