Chest X-ray synthetic data for better testing and evaluation of ML models

Elsa Bismuth
ICME
Stanford University
elsabis@stanford.edu

Alexis Geslin
MSE
Stanford University
geslina@stanford.edu

Magdalini Paschali [*]
Department of Radiology
Stanford University
paschali@stanford.edu

Abstract

The use of Machine Learning (ML) models in radiology has significantly advanced medical imaging diagnostics. However, model performance may not accurately reflect certain underrepresented conditions if such conditions or minorities are rare in the test set. This study demonstrates that synthetic chest X-ray data generated by the state-of-the-art Roentgen model is representative of real data and can address these shortcomings. We generated 600 synthetic images and conducted a comparative analysis with real images using PCA, t-SNE, and histograms, examining both the images and their embeddings through a DenseNet model. Our results indicate that synthetic and real data share similar distributions and feature embeddings. Performance metrics, including AUC, accuracy, and precision, were computed to evaluate a disease classification model from the TorchXrayVision package on both real and synthetic datasets. The results show that synthetic data can enhance model evaluation by providing a more balanced and diverse test set, supporting robust testing for underrepresented diseases. Using synthetic data thus has the potential to ensure fairer, more comprehensive assessments of ML models in radiology.

1. Introduction

The application of Machine Learning (ML) models in biomedical sciences, particularly in radiology, has significantly advanced medical diagnostics based on imaging [1]. In particular, Chest X-Ray (CXR) imaging is a crucial diagnostic tool for various diseases, including cardiomegaly, atelectasis, pleural effusion, and pneumothorax. However, the complexity of such applications requires deep ML models with a large number of parameters, which require extensive and diverse datasets for effective training and evaluation. While large clinical trials can provide reasonable datasets, using data augmentation is a commonly accepted practice to extend the size of datasets and facilitate model training.

However, whether synthetic or real, such datasets may not always be diverse enough and accurately capture underrepresented subgroups. While data augmentation techniques are extensively studied for training, only a few works have explored the use of synthetic imaging as test data [2, 3].

In this work, we investigate if a state-of-the-art chest X-ray synthetic data generative model can be used to create test sets that account for underrepresented conditions or subgroups in the data to better assess the predictive performance of ML models [2]. By incorporating high-fidelity synthetic test data, we aim to determine if the claimed performance of a predictor model, such as a disease detection computer vision model, can be maintained across a wider variety of test samples. This approach addresses the critical issue of dataset diversity and representation, ensuring more robust and reliable predictive models in healthcare.

First, we used a state-of-the-art text-to-image model *Roentgen* [4] to generate chest X-ray synthetic data from text prompts inspired by radiology reports. Second, we used two existing real chest X-ray datasets, namely MIMIC [5, 6] and CheXpert [7] to evaluate the fidelity and diversity of the synthetic data, using techniques such as Principal Component Analysis (PCA), t-Distributed Stochastic Neighbor Embedding (t-SNE) and pixel histograms. Third, we used a state-of-the-art disease classification model from the *TorchXrayVision* package [8] taking X-ray images as inputs and outputting disease scores. One the one hand, we used this model to compare spatial and final embeddings of real and synthetic images to uncover any distribution shifts. On the other hand, we used this model to make disease predictions on both real and synthetic datasets to analyze and conclude if synthetic data can be used to assess the ability of a model to account for underrepresented minorities.

2. Related Work

The efficacy of deep learning models in medical diagnostics has been well-documented, highlighting their potential to enhance diagnostic accuracy and efficiency in clini-

[*] Project advisor, not enrolled in CS231N.

cal settings [1]. Chest X-rays, which are frequently used in the medical field, provide critical information for diagnosing various diseases. Consequently, the application of deep learning methods to analyze chest X-rays has been extensively researched [9, 10, 11]. However, these techniques require large amounts of data for effective training, posing a significant challenge.

To tackle the scarcity of medical X-ray images for deep learning, several approaches have been undertaken. Monshi *et al.* [12] used data augmentation techniques based on segment extraction to enhance their dataset. Others, like Devnath *et al.* used Generative Adversarial Networks (GANs) and transfer learning techniques to augment their training set. GANs have been used for multiple X-ray applications, including abnormality classification [13], disease classification [14], or inpainting [15]. However, most approaches discussed above focus on enhancing a train set to better train a deep learning model and obtain better performance, but not to augment the test sets.

Recent work has explored the use of synthetic data to enhance test sets. Vanbreugel *et al.* [2] focused on improving test sets to better represent underrepresented subgroups and manage distribution shifts. Ktena *et al.* [3] used synthetic data to test models on radiology images, training generative models on the CheXpert dataset to account for minority groups and distribution variations in model evaluation. Coyner *et al.* [16] employed progressively growing generative adversarial networks (PGANs) to generate synthetic retinal vessel maps for robust and privacy-preserving AI model training in retinopathy of prematurity diagnosis. However, none of these generative approaches allow for the synthesis of representative X-ray images from simple text descriptions in a scalable way.

In 2022, Chambon *et al.* tackled that issue and introduced the *Roentgen* model, which generates high-fidelity chest X-ray images from text prompts, effectively addressing the lack of sufficient labeled data in radiology [4].

Our approach builds on these foundations by focusing on using synthetic data to enhance test sets in chest X-ray imaging. Unlike previous work that primarily addresses training data augmentation, we emphasize the importance of diverse and representative test data for robust model evaluation. Leveraging state-of-the-art synthetic data generation techniques, our work aims to ensure that predictive models are thoroughly assessed across a wide range of scenarios.

3. Data

In this work, we used two real chest X-ray datasets. First, we used the MIMIC dataset [5, 6], which is the largest available chest X-ray dataset and the dataset on which the *Roentgen* model has been trained on. The MIMIC dataset contains a total of 377,110 images and associated free-form radiology reports. For this study, we focus on frontal chest X-ray scans, which represent the majority of the dataset. The MIMIC data is available online after taking privacy training (to access patients' health data).

Second, we used the publicly-available CheXpert dataset [7] to be able to compare to a dataset that is not the original dataset used for *Roentgen* training. The CheXpert dataset covers the same diseases as MIMIC, with various degrees of severity, allowing us to compare synthetic images of more severe cases.

We filtered the MIMIC and CheXpert test sets down to patients with the diseases of interest and with frontal chest X-ray images (see section 4.1.2).

Lastly, we compare chest X-ray images from these two real datasets to synthetic images from the dataset we created using *Roentgen*. The methodology to generate such a dataset is further developed in the subsequent section.

Before being forwarded into the *TorchXrayVision* model, both real and synthetic images were preprocessed (resized to be Nx1x224x224 where N is the batch size, normalized and cropped) using *TorchXrayVision* package functions [8].

4. Methods

4.1. Synthetic Data Generation with RoentGen

The *RoentGen* model is an advanced vision-language foundation model adapted specifically for generating synthetic chest X-ray (CXR) images. It builds upon the Stable Diffusion (SD) architecture, leveraging a latent diffusion model (LDM) [17] that combines the capabilities of variational autoencoders (VAEs) and denoising U-Nets. The model is trained using a corpus of publicly available CXR images and their corresponding radiology reports from the MIMIC dataset. *RoentGen* utilizes a text encoder to process free-form medical text prompts and generate high-fidelity, anatomically realistic CXR images.

4.1.1 Model Architecture

RoentGen's architecture is represented in Figure 1 and consists of three main components:

- **Text Encoder**: Converts radiology reports into 768-dimensional embeddings. This encoder is fine-tuned or replaced with domain-specific models like RadBERT [18] or SapBERT [19] for better performance.
- **Conditional Denoising U-Net**: Iteratively denoises random Gaussian noise, conditioned on text embeddings.
- **Variational Autoencoder (VAE)**: Compresses high-dimensional CXR images into lower-dimensional latent representations and decodes them back to the pixel space.

During training, *RoentGen* combines these components to generate high-fidelity synthetic CXR images from text prompts.

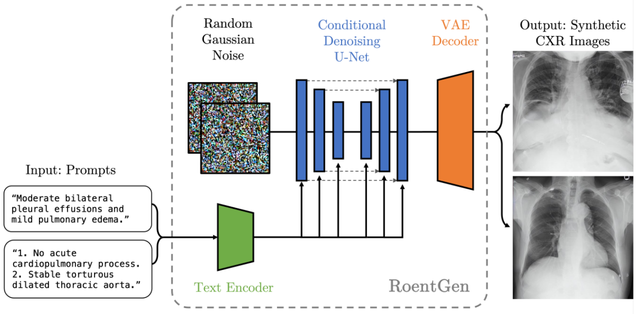

Random
Gaussian
Noise
Conditional
Denoising
U-Net
VAE
Decoder
Output: Synthetic
CXR Images
Input: Prompts
“Moderate bilateral pleural effusions and mild pulmonary edema.”
“1. No acute cardiopulmonary process. 2. Stable torturous dilated thoracic aorta.”
Text Encoder
RoentGen

Figure 1: *RoentGen* Architecture. Source [4]. The model is trained on X-ray images along with the associated radiology text reports from the MIMIC dataset. The core of this stable diffusion model is a VAE, a conditional denoising U-net and a conditioning mechanism (or text encoder).

4.1.2 Data Generation

We focused on four diseases for synthetic data generation: Cardiomegaly, Pleural Effusion, Pneumothorax, and Atelectasis. We also added a ”No Findings” category for a total of 4 conditions and one healthy control category. We extracted reports from the MIMIC test set and created 40 new prompts for each disease, representative of the realistic reports associated with the corresponding disease. Each set of 40 prompts is expanded to a set of 120 prompts by varying the degree of severity of the disease. As a result, a total of 600 prompts were gerenated. An example of short prompt used is: ”*Mild left apical pneumothorax. The right lung appears relatively clear. No pneumothorax on the right side is seen.*”. Using these prompts, we used *RoentGen* via a HuggingFace API token to generate 600 synthetic chest X-ray images. Each image was visually inspected for anatomical relevance, and 14.8% of the images were re-generated to ensure high fidelity.

4.2. Real and synthetic images comparison analysis

4.2.1 Raw Images

To evaluate the similarity between real and synthetic images, we employed the following techniques:

- **Principal Component Analysis (PCA)**: A linear dimensionality reduction technique which captures the principal components explaining the most variance in the data, useful for initial exploratory analysis and identifying broad patterns.
- **t-Distributed Stochastic Neighbor Embedding (t-SNE)**: A non-linear technique for visualizing complex relationships and local structures in 2D and 3D. t-SNE focuses on finding neighboring points, thus detecting clusters, which are not necessarily identified in PCA.
- **Silhouette score on 3D t-SNE**: Evaluates cluster quality. A high score indicates distinct clusters, while a low score indicates overlap. We expect a low score to show synthetic and real images are similar, indicating good representativeness of the synthetic data.
- **Histogram Comparisons**: Provides a straightforward assessment of pixel intensity distributions, highlighting variations in image content.

4.2.2 Image embeddings

To go beyond raw image comparisons, we also compared real and synthetic images embeddings through a DenseNet ML model from *TorchXrayVision* package [8]. DenseNet models introduced by Huang *et al.* in 2016 leverage connectivity between each layer and its subsequent layers to facilitate feature propagation and identification [20]. We used the output of the DenseNet (a spatial embedding of dimension Nx1024x16x16 where N is the batch size) to first visually compare them (cf SI Figure 5). Most importantly, these embeddings were then fed into a ReLu layer and average pooling layer to generate flattened embeddings of dimensions Nx1024). We then applied similar techniques (PCA and t-SNE) to analyze how representative of the real data the synthetic data was. The cosine similarity metric was also used to compare real versus synthetic image embeddings.

4.3. Model Evaluation

Finally, we evaluated the model performance on both real and synthetic datasets using the *TorchXrayVision* package [8]. The $model\_all$ was chosen for its robust training on diverse datasets, including NIH [21], CheXpert, and MIMIC-CXR, which allows it to generalize well across various populations and imaging conditions. This broad training base is essential for assessing performance on synthetic data, ensuring that the evaluation accurately reflects the model’s ability to maintain high performance across different and potentially underrepresented test samples. The core of the model is a DenseNet. The output of the DenseNet is flattened and fed into a Linear classifying layer followed by a Sigmoid function for the final multi-label classification that covers 14 diseases. As a result, the final output is disease probabilities, of size (Nx14).

4.3.1 Test Set Composition

The composition of the real and synthetic test sets for each disease is detailed in Table 1.

4.3.2 Performance Metrics

The chosen metrics—AUC, accuracy (optimal threshold), and precision (optimal threshold)—are particularly relevant to our problem statement. The optimal threshold is computed by finding the point on the ROC curve that maximizes

Disease	CheXpert	Synthetic	MIMIC
Cardiomegaly	175	120	43
Pleural Effusion	120	120	41
Atelectasis	178	120	38
Pneumothorax	10	120	11
“No Findings”	185	120	19
Total	**668**	**600**	**107**

Table 1: Test set composition for each disease across CheXpert, Synthetic, and MIMIC datasets.

the true positive rate while minimizing the false positive rate, essentially where the sum of sensitivity and specificity is highest.

- **AUC** (Area Under the Curve) provides a robust measure of the model’s ability to generalize across different data distributions. A higher AUC on synthetic data suggests that these images effectively capture the variability and complexity of real-world scenarios, which is crucial for robust model evaluation.
- **Accuracy (with optimal threshold)** shows the best-case performance of the model on synthetic versus real data. This metric helps us determine whether synthetic data aids in achieving higher reliability in predictions, which is critical for clinical applications where accurate diagnosis is paramount.
- **Precision (with optimal threshold)** measures the proportion of true positive predictions among all positive predictions at the optimal threshold. High precision is essential in medical diagnostics to reduce false positives, which can lead to unnecessary treatments and patient anxiety. High precision on synthetic data indicates that the synthetic images are realistic enough to challenge the model effectively, validating the use of synthetic data in enhancing test sets.

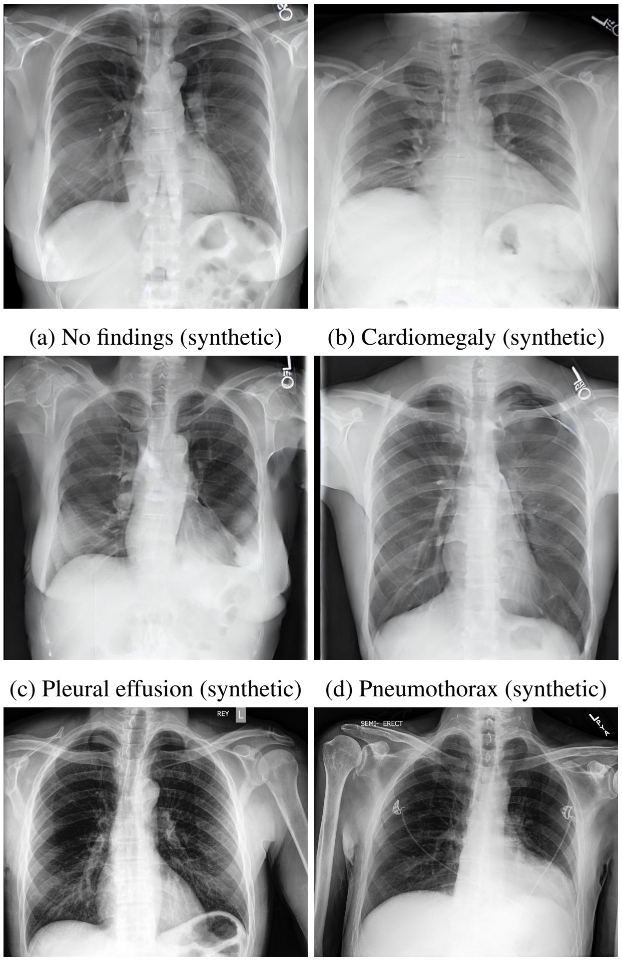
These metrics are relevant because they provide a comprehensive evaluation of model performance, ensuring it can generalize across diverse datasets and maintain reliability in clinical settings.

5. Results and discussion

5.1. Synthetic chest X-ray generation

First, we generated synthetic X-ray images for each of the four diseases and the healthy patients (“No findings”. A few chest X-ray examples are shown in figure 2 (a-d), and two real X-ray images are added (e-f) for comparison.

Given the underrepresentation of conditions like pneumothorax (1.4% of CheXpert test patients), enhancing the test set to include these conditions allows for better model evaluation as we will develop in the following subsections.



(a) No findings (synthetic) (b) Cardiomegaly (synthetic)

(c) Pleural effusion (synthetic) (d) Pneumothorax (synthetic)

(e) No findings (real) (f) Cardiomegaly (real)

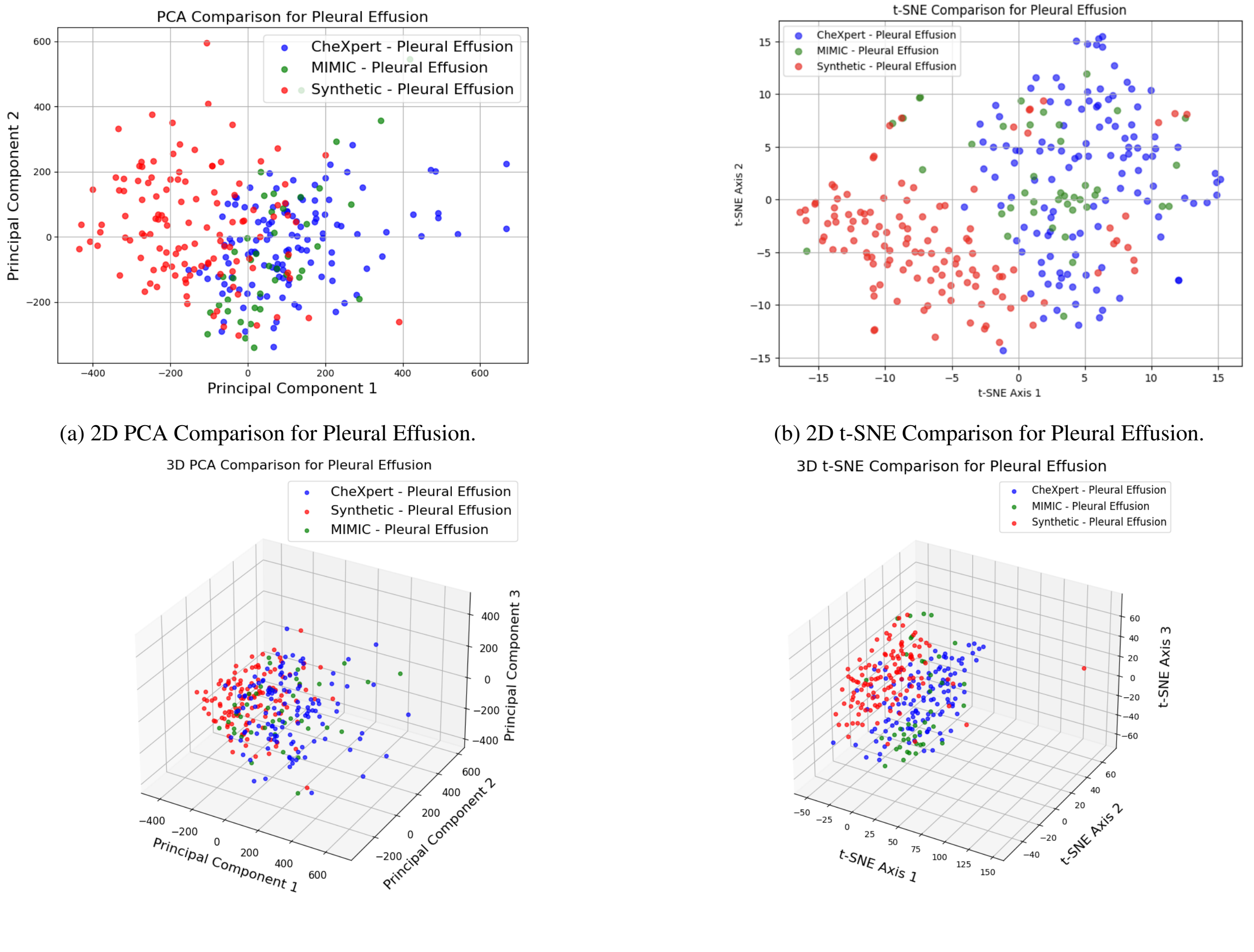
Figure 2: Synthetic chest X-ray images generated with *Roentgen* model

5.2. Comparison between real and synthetic images

5.2.1 Raw images

Our analysis indicates that while synthetic images generally resemble real images, notable differences exist.

- **PCA Plots**: The 2D and 3D PCA plots (e.g., Figures 3a and 3c for Pleural Effusion) reveal overlapping but separate clusters for real and synthetic images. Some variance along the first principal component is observed, explained by different sources. Despite this, there is good overlap in the second and third directions, indicating synthetic and real images are quite similar.
- **t-SNE Plots**: The 2D and 3D t-SNE plots (Figures 3b and 3d) show distinct clusters for real and synthetic images, but the points are still scattered across all three dimensions. This scattering suggests synthetic data can represent real images.



(a) 2D PCA Comparison for Pleural Effusion.

(b) 2D t-SNE Comparison for Pleural Effusion.

(c) 3D PCA Comparison for Pleural Effusion.

(d) 3D t-SNE Comparison for Pleural Effusion.

Figure 3: Comparative Analysis of Real and Synthetic Images. In both PCA and t-SNE representations, synthetic data is spread out in all dimensions, illustrating that synthetic images and real images are not from distinct distributions.

- **Silhouette Scores**: The silhouette scores for Cardiomegaly, Pleural Effusion, Atelectasis, Pneumothorax, and ”No Findings” are 0.145, 0.171, 0.199, 0.066, and 0.140, respectively. These low scores indicate overlapping clusters, demonstrating that synthetic and real images are similar.
- **Histograms**: Pixel intensity histograms on Figure 4 show that both MIMIC and CheXpert real datasets present a uniform distribution with a peak at 0, indicative of the black pixels of the background. However, the synthetic dataset’s pixel distribution is less uniform and slightly shifted toward white pixel values. This may be due to the model accentuating severe disease traits or artifacts in synthetic data.

Our analysis of raw images shows that synthetic images generated by *RoentGen* resemble real images, with some differences in pixel intensity distributions and local data structures. Despite these differences, the synthetic images are sufficiently similar to real ones to augment test sets, providing comprehensive model evaluation. The low silhouette scores further confirm the similarity between synthetic and real images, validating the effectiveness of synthetic data for robust ML model evaluation.

5.2.2 Last Embedded Layer

In addition to investigating the variance and data clustering in the raw images, we also looked at the distributions of real and synthetic image embeddings using the DenseNet model described in section 4.

First, a visual inspection of the spatial embeddings shows that the features extracted by the model are focusing on the same region of the images. Figure SI 5 shows

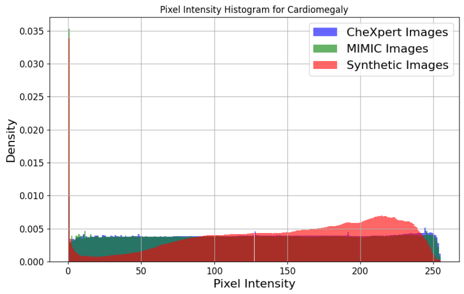


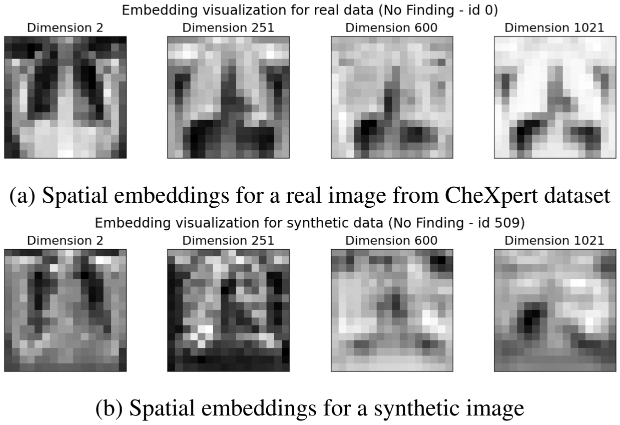
Figure 4: Pixel Intensity Histogram for Cardiomegaly. NB: MIMIC (green) and CheXpert (blue) distribution overlays almost exactly.

4 out of the 1024 spatial embedding dimensions for a real image (from CheXpert dataset) and a synthetic image, both with “No findings”. We can clearly see that the features extracted are looking for specific areas of the chest X-rays, for example, in the first dimension represented on the far left, the features focus on the regions outside the lungs or the rib cage. Notably, these features are looking for the same regions in both the real (Figure SI 5-a) and the synthetic (Figure SI 5-b) images.

Second, to confirm the ability of the *Roentgen* model to generate representative data with high fidelity, we generated synthetic data using the exact text in the radiology report of the real data as a prompt. Figure SI 6 shows both the images and their embeddings. Qualitatively, while the synthetic data is not a exact copy of the real X-ray image, it visually captures the principal elements of it and the embeddings are similar. Quantitatively, the average cosine similary metric across all 1024 embedding dimensions was 76.2% with a standard deviation of 20.1%. This analysis was repeated on three samples, everytime with similar results, confirming that synthetic data can capture real images features.

Finally, Figure 7 represents the first and second dimensions of the entire data (2 real datasets and the synthetic dataset) after applying PCA (7a) and t-SNE (7b). Very interestingly, there is a very good overlap in both the first and second directions between the synthetic data cluster and the real data clusters in PCA. This indicates that the synthetic data embeddings are representative of real image embeddings and do not differ significantly from real data. However, when looking at the t-SNE graph, similarly to the case of raw image t-SNE analysis, we see neat and separate clusters between real and synthetic data along the second dimension. However, importantly, real and synthetic distributions along the first dimension of the t-SNE are very similar.

These embedding results consolidate our conclusion that synthetic data, although not perfectly identical, are representative of real data and captures the relevant features. As a result, synthetic data can be used to enhance test sets.


Embedding visualization for real data (No Finding - id 0)
Dimension 2
Dimension 251
Dimension 600
Dimension 1021
Embedding visualization for synthetic data (No Finding - id 509)
Dimension 2
Dimension 251
Dimension 600
Dimension 1021

(a) Spatial embeddings for a real image from CheXpert dataset

(b) Spatial embeddings for a synthetic image

Figure 5: Real (CheXpert) (a) and synthetic (b) image spatial embeddings through a DenseNet model. Only 4 out of 1024 dimensions are represented. We can visually see that the embeddings are focusing on the same region of the data.

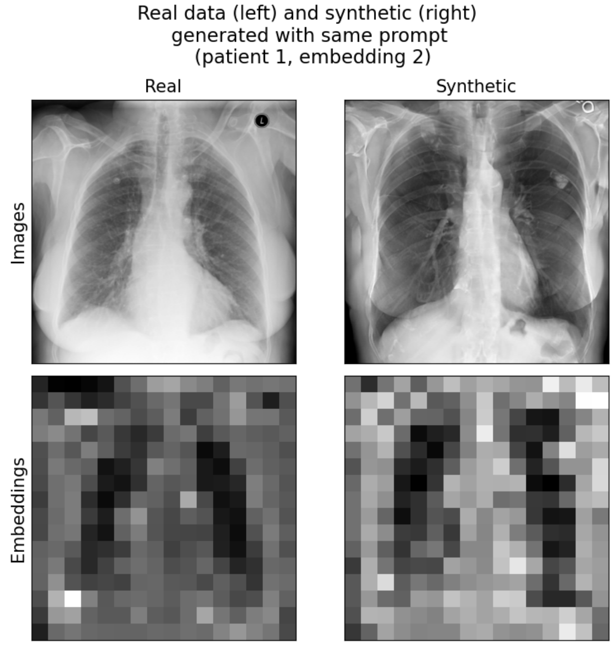
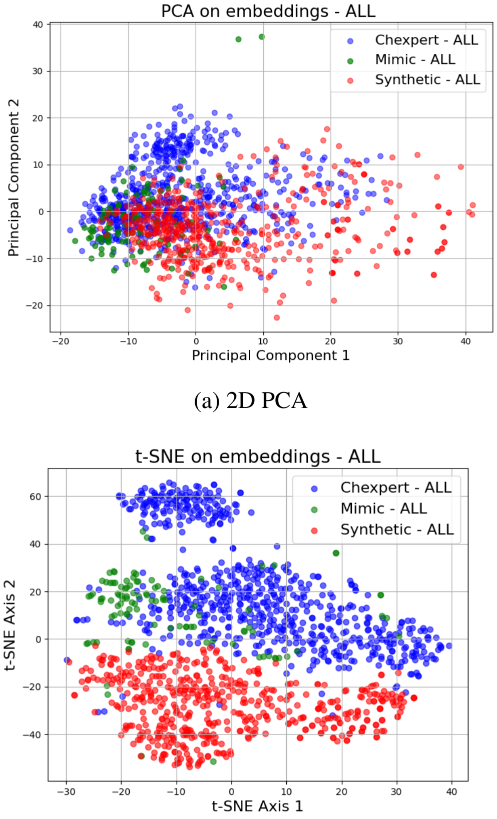

Real data (left) and synthetic (right)
generated with same prompt
(patient 1, embedding 2)
Real
Synthetic
Images
Embeddings

Figure 6: Comparison of a real image from the MIMIC dataset (left) and a synthetic image (right) generated using the exact text from the real image’s radiology report. Embeddings are also shown. The exact text prompt is: *”No new focal consolidation is seen. Mild right apical pleural thickening is seen. No pleural effusion or pneumothorax is seen. The cardiac and mediastinal silhouettes are stable. Small calcification projecting over the upper chest seen on the lateral view is present since at least, and may relate to aortic calcification.”*



(a) 2D PCA

(b) 2D t-SNE

Figure 7: Comparison of embeddings for real and synthetic data. PCA clusters overlap, illustrating that all datasets contribute to explaining some variability, particularly in the second direction. While t-SNE clusters are more distinct, both real and synthetic datasets share similar embeddings along the first dimension.

5.3. Assessment of Prediction Performance

Results in Table 2 show that the *TorchXrayVision* model tends to classify synthetic images slightly more accurately than real images. The relevant metrics are AUC, accuracy (optimal threshold), and precision (optimal threshold). These metrics provide a comprehensive view of the model's ability to discriminate between classes and its performance at the most effective decision threshold.

The results in Table 2 show that the model's AUC and accuracy are similar or slightly higher on synthetic data for all diseases. This slight improvement may be explained by the fact that *Roentgen* model may amplifies disease traits in synthetic data, making classification more consistent. It is worth noting that the precision on synthetic data is generally higher for real data (except for Pneumothorax), indicating that synthetic data can more easily be misclassified positively. These results show that synthetic data performance metrics are in line with real data. Thus, synthetic data can be used to enhance test sets, in particular to account for under-represented minorities.

We use "Pneumothorax" disease as an example use case to illustrate the value of augmenting the test set. The CheXpert and MIMIC real test sets used here only have 10 and 11 positive samples of pneumothorax respectively. As a result, performance metrics of the model on this disease cannot be trusted as much as the others. Using synthetic data with 120 positive pneumothorax cases as a reference, we can have a better assessment of the predictive power of the model for this minority group. From Table 2, we can see that the AUC on synthetic data is about 71.4%, compared to 63.8% and 78.0% for CheXpert and MIMIC respectively. On the one hand, while the CheXpert AUC and accuracy seem in line with other diseases, its precision is off (only 5.2% versus 36.0% for the synthetic data), indicating that the model is not able to pick up positive cases of pneumothorax, despite displaying a good AUC and accuracy. On the other hand, MIMIC exhibits very high AUC and accuracy performance, higher than the synthetic data. However, from diseases with large number of positive samples, performance on synthetic data is expected to be better. This could be an illustration of the lack of pneumothorax samples in the MIMIC dataset leading to artificially inflated results seen here. This "Pneumothorax" use case highlights the benefit of synthetic data in creating balanced test sets.

These findings indicate that high-fidelity synthetic test data can improve model evaluation by providing a comprehensive and diverse test set. This supports the use of synthetic data for fair and robust testing, especially for underrepresented diseases.

Disease	**Data**	**AUC [%]**	**Accuracy [%]**	**Precision [%]**
Cardio.	Synthetic	**88.96**	**82.71**	61.94
Cardio.	CheXpert	72.72	68.61	**71.20**
Cardio.	MIMIC	72.97	72.73	70.21
P.E.	Synthetic	**79.34**	**80.63**	61.16
P.E.	CheXpert	75.05	71.20	62.81
P.E.	MIMIC	70.11	67.05	**77.27**
Atelect.	Synthetic	**73.21**	**66.04**	41.15
Atelect.	CheXpert	61.13	60.52	**67.28**
Atelect.	MIMIC	55.21	56.82	50.00
Pneumo.	Synthetic	71.39	57.71	**36.03**
Pneumo.	CheXpert	63.81	52.10	5.19
Pneumo.	MIMIC	**78.04**	**71.59**	28.13

Table 2: Performance metrics in percentage of the *TorchXrayVision* model on synthetic, CheXpert, and MIMIC datasets. Diseases: Cardio. (Cardiomegaly), P.E. (Pleural Effusion), Atelect. (Atelectasis), Pneumo. (Pneumothorax).

6. Conclusion

This study demonstrates the effectiveness of using synthetic chest X-ray data generated by the *RoentGen* model to enhance test sets for ML models evaluation in radiology. First, we generated a total of 600 synthetic chest X-ray images spanning 5 different conditions. Second, we found that synthetic X-ray images, while not strictly identical, are representative of real X-ray images, with similar feature distributions in PCA and t-SNE for both the images themselves and their embeddings via a DenseNet model. Third, our analysis showed that, using a state-of-the-art disease classification model from chest X-ray *TorchXrayVision*, synthetic data have similar if not slightly higher model evaluation metrics compared to real data. As a result, synthetic data can be used to enhance test sets for a fair model evaluation. As a final use case, we investigated the model performance on “Pneumothorax”, for which both real datasets have limited positive samples. We highlight the challenge of evaluating underrepresented conditions, underscoring the value of synthetic data in such cases.

Future work should focus on incorporating metadata to compare images within specific groups or categories of people could enhance the evaluation of model performance across different demographics. Expanding the use of synthetic data to other imaging modalities and conditions could also broaden its applicability, ensuring comprehensive and equitable evaluations of ML models across various medical fields.

7. Contributions & Acknowledgements

Elsa and Alexis contributed equally to the work. More specifically, Alexis worked on synthetic data generation, MIMIC dataset pre-processing and embeddings analysis while Elsa worked on raw image analysis and ML prediction performance. The project was proposed and mentored by Magdalini, who is not enrolled in CS231N. Magdalini also provided access to the *Roentgen* model on HuggingFace.

Elsa and Alexis would like to particularly thank Magdalini for her support and for providing insights throughout the project.

As specified in the *Methods*, we used the *Roentgen* model via the Huggingface API (huggingface.co) and the *TorchXrayVision* package available for download (github.com). This project is not shared with another class.

References

[1] Andre Esteva, Katherine Chou, Serena Yeung, Nikhil Naik, Ali Madani, Ali Mottaghi, Yun Liu, Eric Topol, Jeff Dean, and Richard Socher. Deep learning-enabled medical computer vision. *npj Digital Medicine*, 4(1):5, January 2021.

[2] Boris van Breugel, Nabeel Seedat, Fergus Imrie, and Mihaela van der Schaar. Can you rely on your model evaluation? improving model evaluation with synthetic test data. In *Advances in Neural Information Processing Systems*, volume 36, pages 1889–1904. Curran Associates, Inc., 2023.

[3] I. Ktena, O. Wiles, I. Albuquerque, et al. Generative models improve fairness of medical classifiers under distribution shifts. *Nature Medicine*, 30:1166–1173, 2024.

[4] Pierre Chambon, Christian Bluethgen, Jean-Benoit Delbrouck, Rogier Van der Sluijs, Małgorzata Połacin, Juan Manuel Zambrano Chaves, Tanishq Mathew Abraham, Shivanshu Purohit, Curtis P. Langlotz, and Akshay Chaudhari. Roentgen: Vision-language foundation model for chest x-ray generation. *arXiv preprint*, arXiv:2211.12737, 2022.

[5] Alistair E. W. Johnson, Tom J. Pollard, Nathaniel R. Greenbaum, Matthew P. Lungren, Chih ying Deng, Yifan Peng, Zhiyong Lu, Roger G. Mark, Seth J. Berkowitz, and Steven Horng. Mimic-cxr-jpg, a large publicly available database of labeled chest radiographs. *arXiv preprint*, arXiv:1901.07042, 2019.

[6] Alistair E W Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-Ying Deng, Roger G Mark, and Steven Horng. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific Data*, 6(1):317, December 2019.

[7] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, Jayne Seekins, David A. Mong, Safwan S. Halabi, Jesse K. Sandberg, Ricky Jones, David B. Larson, Curtis P. Langlotz, Bhavik N. Patel, Matthew P. Lungren, and Andrew Y. Ng. Chexpert: a large chest radiograph dataset with uncertainty labels and expert comparison. *Proceedings of the AAAI conference on artificial intelligence*, 33(01):590–597, 2019.

[8] Joseph Paul Cohen, Joseph D. Viviano, Paul Bertin, Paul Morrison, Parsa Torabian, Matteo Guarrera, Matthew P Lungren, Akshay Chaudhari, Rupert Brooks, Mohammad Hashir, and Hadrien Bertrand. Torchxrayvision: A library of chest x-ray datasets and models. In *Proceedings of The 5th International Conference on Medical Imaging with Deep Learning*, volume 172 of *Proceedings of Machine Learning Research*, pages 231–249. PMLR, 06–08 Jul 2022.

[9] Dulani Meedeniya, Hashara Kumarasinghe, Shammi Kolonne, Chamodi Fernando, Isabel De la Torre Díez, and Gonçalo Marques. Chest X-ray analysis empowered with deep learning: A systematic review. *Applied Soft Computing*, 126:109319, 2022.

[10] Stefanus Tao Hwa Kieu, Abdullah Bade, Mohd Hanafi Ahmad Hijazi, and Hoshang Kolivand. A survey of deep learning for lung disease detection on medical images: State-of-the-art, taxonomy, issues and future directions. *Journal of Imaging*, 6(12), 2020.

[11] Erdi Çallı, Ecem Sogancioglu, Bram van Ginneken, Kicky G. van Leeuwen, and Keelin Murphy. Deep learning

for chest X-ray analysis: A survey. *Medical Image Analysis*, 72:102125, 2021.

[12] Maram Mahmoud A. Monshi, Josiah Poon, Vera Chung, and Fahad Mahmoud Monshi. Covidxraynet: Optimizing data augmentation and cnn hyperparameters for improved covid-19 detection from cxr. *Computers in Biology and Medicine*, 133:104375, 2021.

[13] Ali Madani, Mehdi Moradi, Alexandros Karargyris, and Tanveer Syeda-Mahmood. Chest X-ray generation and data augmentation for cardiovascular abnormality classification. In Elsa D. Angelini and Bennett A. Landman, editors, *Medical Imaging 2018: Image Processing*, volume 10574, page 105741M. International Society for Optics and Photonics, SPIE, 2018.

[14] Devansh Srivastav, Akansha Bajpai, and Prakash Srivastava. Improved classification for pneumonia detection using transfer learning with gan based synthetic image augmentation. In *2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*, pages 433–437, 2021.

[15] Ecem Sogancioglu, Shi Hu, Davide Belli, and Bram van Ginneken. Chest X-ray inpainting with deep generative models. *arXiv preprint*, arXiv:1809.01471, 2018.

[16] Aaron S Coyner, Jimmy S Chen, Ken Chang, Praveer Singh, Susan Ostmo, R V Paul Chan, Michael F Chiang, Jayashree Kalpathy-Cramer, and J Peter Campbell. Synthetic medical images for robust, privacy-preserving training of artificial intelligence: Application to retinopathy of prematurity diagnosis. *Ophthalmology Science*, 2(2):100126, 2022.

[17] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, June 2022.

[18] An Yan, Julian McAuley, Xing Lu, Jiang Du, Eric Y. Chang, Amilcare Gentili, and Chun-Nan Hsu. RadBERT: Adapting transformer-based language models to radiology. *Radiology: Artificial Intelligence*, 4(4):e210258, 2022. PMID: 35923376.

[19] Fangyu Liu, Ehsan Shareghi, Zaiqiao Meng, Marco Basaldella, and Nigel Collier. Self-alignment pretraining for biomedical entity representations. In *North American Chapter of the Association for Computational Linguistics*, 2020.

[20] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger. Densely connected convolutional networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2261–2269, Los Alamitos, CA, USA, jul 2017. IEEE Computer Society.

[21] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M. Summers. 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)*. IEEE, July 2017.