Medical Image Synthesis for Data Augmentation and Anonymization via Diffusion

Nashir Janmohamed
Computer Science, BS
nashir@knights.ucf.edu

1. Problem Definition

This research investigates the use of diffusion models [2, 5, 8, 18, 23] to generate high-fidelity medical imaging data. The motivation is twofold:

1. Deep learning models typically require large amounts of data to perform well on a task. Augmenting the size of the training data-set with diffusion-generated synthetic images can potentially improve the performance of a model, which would better assist in a medical setting.

2. Medical practices may not contribute their data to a large centralized data-set due to patient privacy concerns. Pretraining on synthetic images and fine tuning on a small number of real patient images may be sufficient to develop a model that is performant enough to be useful while minimizing use of real patient data.

This work aims to generate high-quality diffusion-generated medical images and compare image classification performance using diffusion—generated synthetic medical images as augmentation vs. no augmentation vs. traditional augmentation. We believe that using diffusion-generated medical images for pre-training will produce a more performant classifier, which has implications for data availability and patient privacy concerns.

The source code is available on GitHub.

2. Related Work

Recent works have shown that diffusion beats GANs on the image synthesis task on ImageNet [2] and on medical imaging data-sets [13].

Many works have investigated the use of GANs to generate medical imagery [25, 24, 17], and in particular, there are a few works that use GANs to generate training data for the classification or segmentation task in medical imaging [16, 20, 19].

As of the time of this writing, fewer works have addressed the use of diffusion to generate medical images [8]. [11] generate diffusion medical images of the Cancer Genome Atlas (TCGA) data-set [21] and show the improvement of the proposed method over GAN-based approaches. [10] proposes a diffusion deformable model (DDM) to generate 3D+t (4D) temporal volume images. [14] use diffusion to generate a data-set of 1000 synthetic images and compare the FID, MS-SSIM, and 4-G-R-SSIM to GAN and VAE approaches. They do not compare the relative performance of classifiers trained on diffusion data. [1] fine-tunes components of the Stable Diffusion [15] pipeline to generate medical images, but does not use the generated data to train classifiers. [9] use DPMs to synthesize high quality 3D medical imaging data (CT and MRI). They also demonstrate self-supervised pre-training on synthetic images can be used in scarce data settings to improve the performance of breast segmentation models.

Relevant source codes for general diffusion [3, 22] and 2D diffusion on medical images exist [12].

3. Technical Approach

To generate diffusion data with labels, we use classifier free guidance [6]. Due to computational constraints, we train the diffusion model to generate downscaled x-ray images. As a preprocessing step, we created a transform to downsize and center crop the original x-ray images from the Chexpert [7] dataset and use these for both 1) training the diffusion model and 2) in the downstream image evaluation and classification task.

An alternative to this approach is to use a super-resolution model to upsample the low resolution generated images before comparing the synthetic imagery to the original scale ground truth images, and using the upscaled diffusion imagery in the classification task. This may be the subject of future work to improve performance.

After downscaling the ground truth, we use the downscaled data to train the diffusion models, and use the downscaled real data when training classifiers. We then 1) compare the relative accuracy of classifiers trained on each of these datasets (with the expectation that the real data will

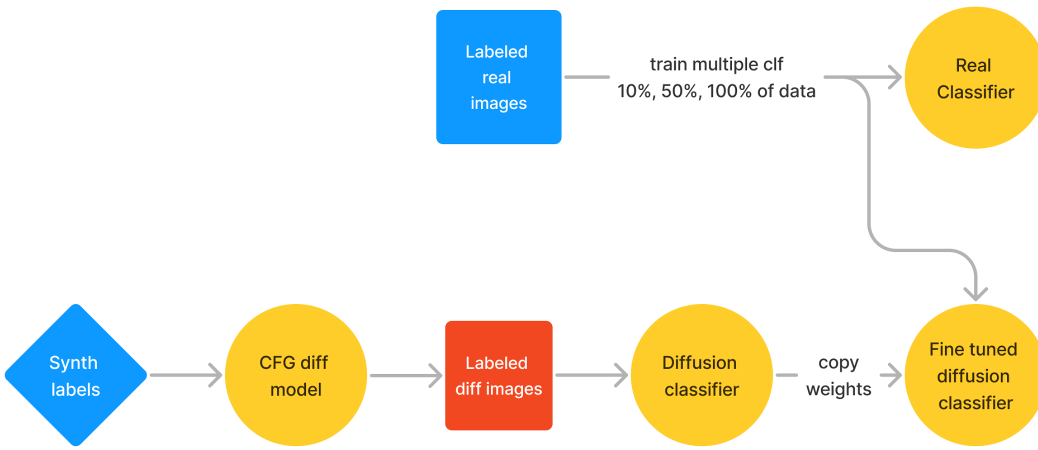

Figure 1: Classifier training pipeline; we have three primary approaches 1) all real data, 2) all synthetic data, 3) pretrain on synthetic data and fine tune on real data. All classifiers are Resnet18, with initial weights from ImageNet1k. We evaluate the performance of these classifiers on a held out test set of real data.

yield the highest accuracy), and 2) compute image quality scores (i.e. FID, SSIM) on the real vs. synthetic datasets to gauge their quality.

As mentioned above, diffusion models have been shown to beat GANs at image synthesis on natural [2] and medical [13] images, and we believe this enhancement in image quality will improve the use of generative models as a data augmentation and privacy preservation tool.

3.1. Diffusion model training

We train state-of-the art classifier-free-guidance diffusion models to generate realistic samples from a 2D medical imaging data-set (CheXpert [7]). We leverage https://github.com/lucidrains/denoising-diffusion-pytorch for training a model to generate chest xray images with class labels.

3.2. Image quality evaluation

Figure 2 shows the process for comparing the image distribution from two datasets; the results are outlined in Table 1.

4. Experiments and results

4.1. Diffusion model training

After preprocessing the input images to be 28x28, we use the U-Net as the diffusion model backbone and the following hyperparameters:

- Optimizer: Adam
- Learning rate = 1e-4
- Beta (optimizer) = (0.9, 0.999)

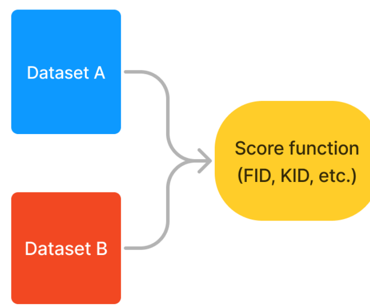

Figure 2: Computation of image quality scores based on the real vs. synthetic datasets constructed using the pipeline shown in Figure 1.

- Eps = 1e-8
- Number of epochs = 300
- Loss: L1 loss
- Beta schedule (diffusion): cosine
- Batch size = 4

We would like to train the model for more epochs, or if more computational resources were available, train a model to generate higher resolution images. As can be seen in Figure 3, the loss didn't plateau, and the model performance could benefit from more epochs. The model was trained on a 4GB GTX 1050 Ti graphics card; on this hardware, each epoch took approximately 1 minute and 5 seconds, and sampling 500 images takes approximately 3 minutes.

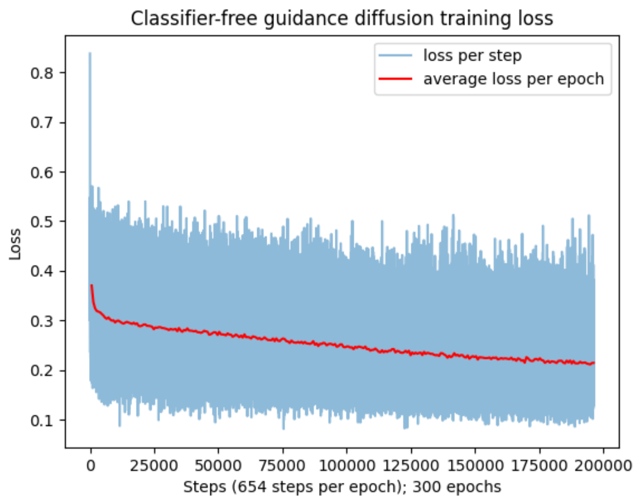

Figure 3: Loss plot from training diffusion model for 300 epochs.

4.2. Image quality metrics

To evaluate the quality of the images we use two metrics. The first is Fréchet inception distance [4], which compares the distribution of generated images with the distribution of a set of real images ("ground truth"). A low FID score is better. The Fréchet distance $d(.,.)$, is computed between the Gaussian with mean $(\mathbf{m}, \mathbf{C})$ obtained from $p(.)$ and the Gaussian with mean $(\mathbf{m_w}, \mathbf{C_w})$ obtained from $p_w(.)$. The Fréchet inception distance (FID) is given by

$$d^2((\mathbf{m}, \mathbf{C}), \mathbf{m_w}, \mathbf{C_w})) = \\ \|\mathbf{m} - \mathbf{m_w}\|_2{}^2 + \mathrm{Tr}(\mathbf{C} + \mathbf{C_w} - 2(\mathbf{C}\mathbf{C_w})^{1/2})$$

where Tr(.) indicates the trace (the sum of elements on the main diagonal).

According to W&B, the minimum suggested sample size for comparisons with FID is 10k; we don't have this many real xray images; instead we compare 1k samples of each type (except for real-test, because there are only 624 test samples).

The other metric we use is the structural similarity metric, defined as follows:

$$\mathrm{SSIM}(x, y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)}$$

where:

- μ_x: average of x
- μ_y: average of y
- σ_x^2: variance of x
- σ_y^2: variance of y
- σ_{xy}: the covariance of x and y
- $c_1 = (k_1L)^2$, $c_2 = (k_2L)^2$: two variables to stabilize the division with weak denominator
- L the dynamic range of the pixel values
- $k_1 = 0.01$ and $k_2 = 0.03$ by default.

The resulting SSIM index ranges from -1 to +1. A value of +1 is achieved only with the complete authenticity of the samples.

Dataset A	Dataset B	FID (↓)	SSIM (↑)
Real train	Real test	31.25	0.2996
Diffusion	Real train	193.922	0.2713
Diffusion	Real test	207.901	0.2833
Real train	FMNIST	176.893	0.0623
Real test	FMNIST	175.255	0.0757
Diffusion	FMNIST	335.008	0.0424

Table 1: Image quality metrics computed on pairs of datasets; we expect the diffusion generated data to be most similar to the real train data, but also similar to the real test data.

4.3. Classifier training

We evaluate the model by using the generated images to augment a training data-set of real (downscaled) chest x-ray images. We train a classifier on all pairs of 1) the original training data, 2) the synthetic training data, 3) the original training data with data augmentations, and 4) the synthetic training data with data augmentations. We analyze the relative performance to determine 1) the efficacy of diffusion as a data augmentation tool and 2) the potential for diffusion as a privacy protecting mechanism.

See Appendix A for training loss/accuracy plots, as well as confusion matrices for each of the items listed in Table 2.

Hyperparameters:

- Optimizer: SGD
- Learning rate = 0.001
- Momentum = 0.9
- Number of epochs = 20
- LR scheduler: Exponential
- Step size = 7

#	Geo aug?	Synthetic PT?	Real data FT%	Weighted loss?	Acc.	Acc. (N)	Acc. (P)	P	R	F1
1	x	x	100	x	0.811	0.538	0.974	0.853	0.753	0.773
2	x	x	100	✓	0.829	0.590	0.972	0.862	0.781	0.798
3	x	✓	100	x	0.811	0.526	**0.982**	0.861	0.754	0.771
4	x	✓	100	✓	0.829	0.585	0.974	0.864	0.780	0.798
5	✓	x	100	x	0.827	0.650	0.933	0.835	0.791	0.804
6	✓	x	100	✓	0.833	0.585	**0.982**	**0.875**	0.784	0.803
7	✓	✓	100	x	**0.845**	0.667	0.951	0.859	**0.809**	**0.824**
8	✓	✓	100	✓	0.824	**0.705**	0.895	0.818	0.800	0.807
9	x	x	10	x	0.779	0.573	0.903	0.779	0.738	0.748
10	x	x	10	✓	0.772	0.530	0.918	0.780	0.724	0.735
11	x	✓	10	x	0.800	0.581	0.931	0.811	0.756	0.769
12	x	✓	10	✓	0.748	0.449	0.928	0.763	0.688	0.697
13	✓	x	10	x	**0.812**	0.581	0.951	0.834	0.766	0.782
14	✓	x	10	✓	0.792	0.603	0.905	0.792	0.754	0.764
15	✓	✓	10	x	0.804	0.538	**0.964**	**0.838**	0.751	0.767
16	✓	✓	10	✓	0.809	**0.628**	0.918	0.813	**0.773**	**0.785**
17	x	✓	0	x	0.729	0.682	0.808	0.730	0.745	0.725
18	✓	✓	0	x	0.662	0.628	0.718	0.662	0.673	0.657

Table 2: Metrics computed for each of the models trained; we examine the data scare and data rich setting to evaluate how the use of pretraining on synthetic diffusion data affects the classification performance. The best performance in each category (0%, 10%, 100% of real data used) is bolded.

- Gamma = 0.1
- Batch size = 32
- Geo aug: whether or not geometric augmentation (random horizontal flip; affine transformation - 30 degrees, 0.1 by 0.1 translation) are used
- Synthetic PT: whether or not the model was pretrained on the diffusion generated imagery
- Real data FT%: How much real data was used for fine-tuning; note that if there was no synthetic pretraining, the fine tuning is essentially just training on the given percentage of the real data. If the fine tune percentage is 0%, then the model was trained only on diffusion-generated images.
- Weighted loss: whether or not a weighted cross entropy (with weights given by label proportions in the dataset) is used.
- Acc: Overall accuracy on held out test set
- Acc (N): Accuracy on xray images without pneumonia
- Acc (P): Accuracy on xray images with pneumonia
- P: Precision = $\frac{TP}{TP+FP}$
- R: Recall = $\frac{TP}{TP+FN}$
- F1: F1 score = $2\frac{PR}{P+R}$

Note: the synthetic data is balanced, so using weighted loss wouldn't make a difference; therefore we only train two models with no finetuning on real data; with and without geometric augmentation.

In the diagnosis setting, false negatives are more harmful than false positives; thus we should choose the model with the highest recall, rather than the highest precision. In both the data scarce (10% fine tuning) and data rich (100% fine tuning) environments, synthetic pretraining produces the model with the highest recall. However, geometric data augmentation alone is still very effective, and much less computationally demanding than our proposed approach. It is possible that this is due to the scale of the data, and using higher resolution diffusion generated images would produce better results.

4.4. Diffusion generated imagery

During training we sample images after every 5 epochs to qualitatively evaluate how the image quality changes over the course of training. Some sample images are shown in Figures [4-7]. Figure 8 shows a sample of the real down-scaled xray images.

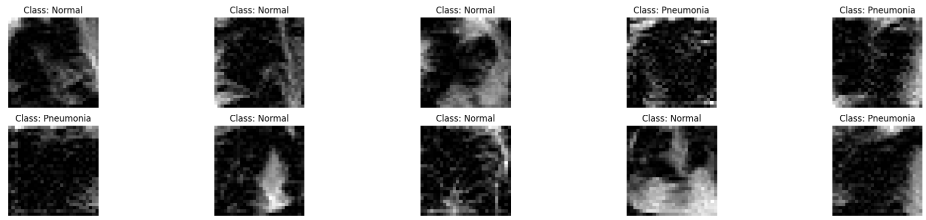

Figure 4: Diffusion generated imagery after 1 epoch of training.

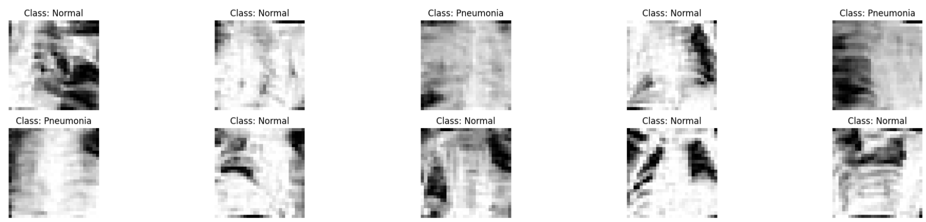

Figure 5: Diffusion generated imagery after 5 epochs of training.

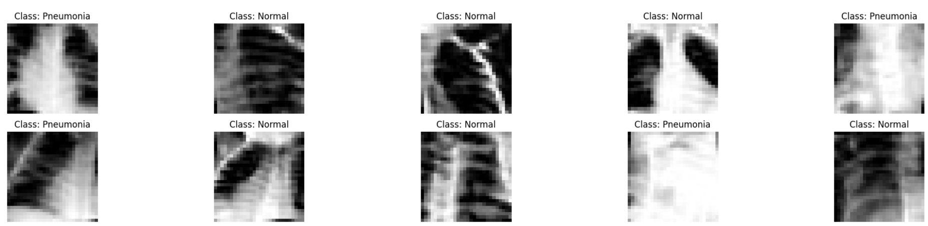

Figure 6: Diffusion generated imagery after 45 epochs of training.

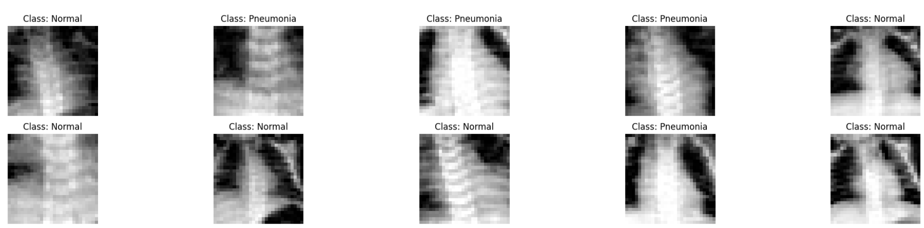

Figure 7: Diffusion generated imagery after 290 epochs of training.

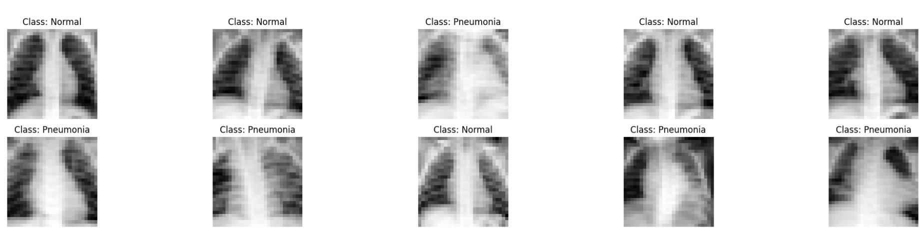

Figure 8: Real downscaled xray imagery.

5. Conclusion

In this work, we examined the use of conditional (classifier-free guidance) diffusion models to generate synthetic x-ray images; we evaluated the generated images based on image quality metrics such as FID, KID, and MS-SSIM, and trained downstream classification models using the synthetically generated data for pretraining/data augmentation. Due to computational constraints, we generated low resolution diffusion imagery, and trained the classifiers on downscaled real data. However, our diffusion pipeline can be easily modified to train a model to generate higher resolution images, and our classification pipeline is set up to accept arbitrary RGB image datasets. Ideally, future work would explore 1) creating diffusion images with higher image resolution, or 2) using a super resolution model to upscale the generated diffusion imagery in order to improve on our results.

We also would like to train the diffusion model for longer, as well as compute the FID, etc. scores every epoch to see how they trend over time.

References

[1] Pierre Chambon, Christian Bluethgen, Curtis P. Langlotz, and Akshay Chaudhari. Adapting Pretrained Vision-Language Foundational Models to Medical Imaging Domains, Oct. 2022. arXiv:2210.04133 [cs]. 1

[2] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794, 2021. 1, 2

[3] Hugging Face. The Annotated Diffusion Model. 1

[4] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium, Jan. 2018. arXiv:1706.08500 [cs, stat]. 3

[5] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. 1

[6] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 1

[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, Jan. 2019. arXiv:1901.07031 [cs, eess]. 1, 2

[8] Amirhossein Kazerouni, Ehsan Khodapanah Aghdam, Moein Heidari, Reza Azad, Mohsen Fayyaz, Ilker Hacihaliloglu, and Dorit Merhof. Diffusion models for medical image analysis: A comprehensive survey. *arXiv preprint arXiv:2211.07804*, 2022. 1

[9] Firas Khader, Gustav Mueller-Franzes, Soroosh Tayebi Arasteh, Tianyu Han, Christoph Haarburger, Maximilian Schulze-Hagen, Philipp Schad, Sandy Engelhardt, Bettina Baessler, Sebastian Foersch, Johannes Stegmaier, Christiane Kuhl, Sven Nebelung, Jakob Nikolas Kather, and Daniel Truhn. Medical Diffusion: Denoising Diffusion Probabilistic Models for 3D Medical Image Generation, Jan. 2023. arXiv:2211.03364 [cs, eess]. 1

[10] Boah Kim and Jong Chul Ye. Diffusion deformable model for 4d temporal medical image generation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part I*, pages 539–548. Springer, 2022. 1

[11] Puria Azadi Moghadam, Sanne Van Dalen, Karina C. Martin, Jochen Lennerz, Stephen Yip, Hossein Farahani, and Ali Bashashati. A morphology focused diffusion probabilistic model for synthesis of histopathology images. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 2000–2009, January 2023. 1

[12] mueller franzes. Medfusion - Medical Denoising Diffusion Probabilistic Model, Feb. 2023. original-date: 2022-12-14T08:15:52Z. 1

[13] Gustav Müller-Franzes, Jan Moritz Niehues, Firas Khader, Soroosh Tayebi Arasteh, Christoph Haarburger, Christiane Kuhl, Tianci Wang, Tianyu Han, Sven Nebelung, Jakob Nikolas Kather, and Daniel Truhn. Diffusion Probabilistic Models beat GANs on Medical Images, Dec. 2022. arXiv:2212.07501 [cs, eess]. 1, 2

[14] Walter HL Pinaya, Petru-Daniel Tudosiu, Jessica Dafflon, Pedro F Da Costa, Virginia Fernandez, Parashkev Nachev, Sebastien Ourselin, and M Jorge Cardoso. Brain imaging generation with latent diffusion models. In *Deep Generative Models: Second MICCAI Workshop, DGM4MICCAI 2022, Held in Conjunction with MICCAI 2022, Singapore, September 22, 2022, Proceedings*, pages 117–126. Springer, 2022. 1

[15] 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*, pages 10684–10695, 2022. 1

[16] Hoo-Chang Shin, Neil A. Tenenholtz, Jameson K. Rogers, Christopher G. Schwarz, Matthew L. Senjem, Jeffrey L. Gunter, Katherine Andriole, and Mark Michalski. Medical Image Synthesis for Data Augmentation and Anonymization using Generative Adversarial Networks, Sept. 2018. arXiv:1807.10225 [cs, stat]. 1

[17] Youssef Skandarani, Pierre-Marc Jodoin, and Alain Lalande. GANs for Medical Image Synthesis: An Empirical Study, July 2021. arXiv:2105.05318 [cs, eess]. 1

[18] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 1

[19] Yi Sun, Peisen Yuan, and Yuming Sun. MM-GAN: 3D MRI Data Augmentation for Medical Image Segmentation via Generative Adversarial Networks. In *2020 IEEE International Conference on Knowledge Graph (ICKG)*, pages 227–234, Aug. 2020. 1

[20] Shobhita Sundaram and Neha Hulkund. GAN-based Data Augmentation for Chest X-ray Classification, July 2021. arXiv:2107.02970 [cs, eess]. 1

[21] Katarzyna Tomczak, Patrycja Czerwińska, and Maciej Wiznerowicz. Review the cancer genome atlas (tcga): an immeasurable source of knowledge. *Contemporary Oncology/Współczesna Onkologia*, 2015(1):68–77, 2015. 1

[22] Phil Wang. lucidrains/denoising-diffusion-pytorch, Feb. 2023. original-date: 2020-08-26T02:22:10Z. 1

[23] Ling Yang, Zhilong Zhang, Yang Song, Shenda Hong, Runsheng Xu, Yue Zhao, Yingxia Shao, Wentao Zhang, Bin Cui, and Ming-Hsuan Yang. Diffusion Models: A Comprehensive Survey of Methods and Applications, Oct. 2022. arXiv:2209.00796 [cs]. 1

[24] Xin Yi. Awesome GAN for Medical Imaging, Feb. 2023. original-date: 2017-08-23T20:38:25Z. 1

[25] Xin Yi, Ekta Walia, and Paul Babyn. Generative adversarial network in medical imaging: A review. *Medical Image Analysis*, 58:101552, Dec. 2019. 1

A. Classifier training plots

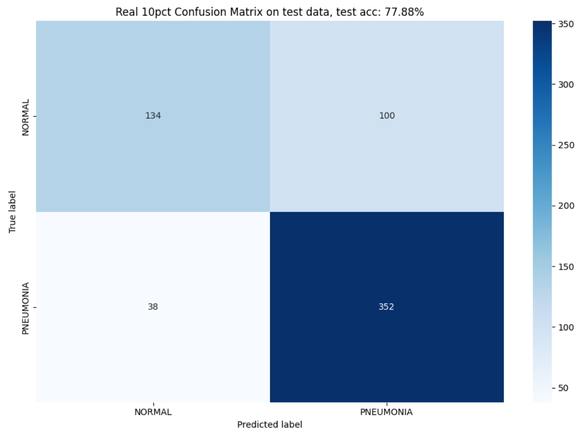

Figure 9: Confusion matrix: No geometric data augmentation, no synthetic pre-training, no weighted loss, 10% real.

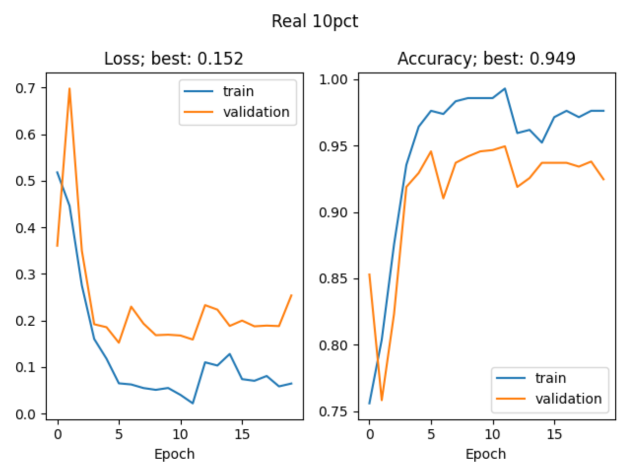

Figure 10: Train metrics: No geometric data augmentation, no synthetic pre-training, no weighted loss, 10% real.

See results folder in project source code for all classifier training and confusion matrix plots.

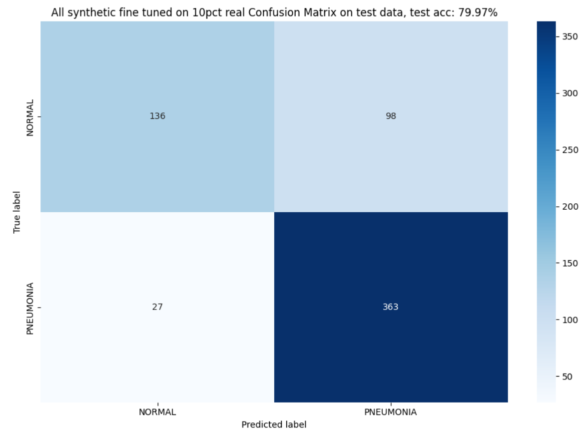

Figure 11: Confusion matrix: No geometric data augmentation, synthetic pre-training, no weighted loss, 10% real.

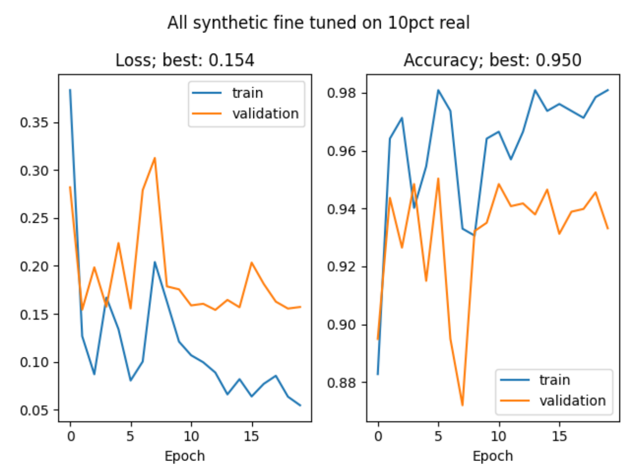

Figure 12: Train metrics: No geometric data augmentation, synthetic pre-training, no weighted loss, 10% real.