# Evaluating Generative Models in Medical Imaging

Liyue Fan
*Department of Computer Science*
*UNC Charlotte*
Charlotte, NC
liyue.fan@charlotte.edu

Ashley Bang*
*UNC Charlotte*
Charlotte, NC
sbang6@charlotte.edu

Luca Bonomi
*Department of Biomedical Informatics*
*Vanderbilt University Medical Center*
Nashville, TN
luca.bonomi@vumc.org

***Abstract*—Data synthesis can address important data availability challenges in biomedical informatics. Quantitative evaluation of generative models may help understand their applications to synthesizing biomedical data. This poster paper examines state-of-the-art generative models used in medical imaging, such as StyleGAN and DDPM models, and evaluates their performance in learning data manifolds and in the visible features of generated samples. Results show that existing generative models have much to improve based on the studied measures.**

***Index Terms*—Medical imaging, generative models, data synthesis**

## I. Introduction

Data synthesis can augment available training data and benefit artificial intelligence (AI) applications, e.g., deep learning [1], [2]. In biomedical domains, data is highly localized; large training data may not be readily available or easily curated due to privacy concerns. Therefore, generating *high-quality* synthetic data is increasingly important for biomedical applications [2], [3].

Medical imaging informatics have been at the forefront of AI technology development and deployment [4]. A number of studies have been conducted to employ state-of-the-art generative models for synthesizing medical images. Notably, generative adversarial networks (GANs) and variants have been applied to a wide range of imaging modalities, such as X-ray and MRI [5]. Recent research adopted denoising diffusion probabilistic models (DDPMs) which demonstrated superior performance to that of GANs [6], [7]. As generative models have developed rapidly and their impacts on medicine/human health are paramount, it is important to understand the current state of medical imaging synthesis.

Compared to existing literature surveys [2], [8], the goal of this preliminary study is to examine state-of-the-art generative models used in medical imaging (such as StyleGAN and DDPM models), through the lens of *interpretable*, *quantitative* performance metrics. Specifically, this study employs a set of classic *spatial features* to contrast real samples with synthetic samples. Furthermore, this study incorporates the improved precision and recall metric to evaluate the *manifolds* learned by those generative models. Our empirical results reveal important gaps in medical imaging synthesis.

*undergraduate student

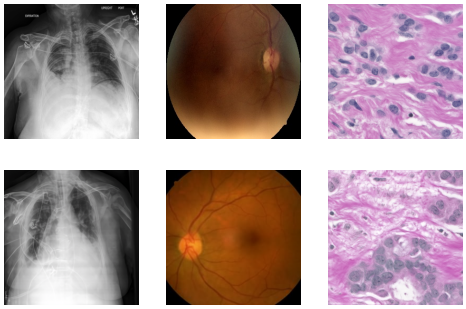


Fig. 1: Sample real (*top*) and generated (*bottom*) images.

## II. Material and Methods

### A. Data Generation

Three publicly available datasets were used in this study. Note that where sample labels are available, we form balanced data to minimize the impacts of disease label distributions. **ChestXray**: 15738 real samples equally distributed in "cardiomegaly" and "no cardiomegaly" classes were taken from the CheXpert dataset [9]. Note that "no cardiomegaly" is the minority class with only 7869 samples. All images were scaled to 256x256. **Fundus**: 6540 real samples equally distributed in "referable glaucoma" and "no referable glaucoma" classes were taken from the AIROGS dataset [10]. Note that "referable glaucoma" is the minority class with only 3270 samples. All images were center cropped and scaled to 256x256. **Histology**: 6000 real samples were taken from the BreCaHAD dataset [11]. Each sample is 512x512 and randomly cropped from 1360x1024 raw images. Same amounts of synthetic samples were generated for all datasets. For ChestXray and Fundus datasets, we generated class-balanced synthetic samples using pre-trained DDPM models from [7], which were trained on CheXpert and AIROGS. For Histology, we generated synthetic samples using a pre-trained StyleGAN2 model from [12], which was trained on BreCaHAD. Samples images are presented in Figure 1.

### B. Interpretable, Quantitative Measures

**Image Spatial Features**. As generated images may differ from real images in visual features, we consider a range of classic spatial features that can be efficiently extracted. The colorfulness index (CFI) [13] approximates the human perception of colorfulness in natural images. Brenne, Tenengrad and Laplacian gradients (BIQ, TIQ, and LIQ) [14] characterize the

TABLE I: Improved precision and recall for generative models.

| Dataset | Precision | Recall | FID |
|---|---|---|---|
| ChestXray | 0.69 | 0.30 | 11.51 |
| Fundus | 0.14 | 0.05 | 34.06 |
| Histology | 0.10 | 0.28 | 17.96 |

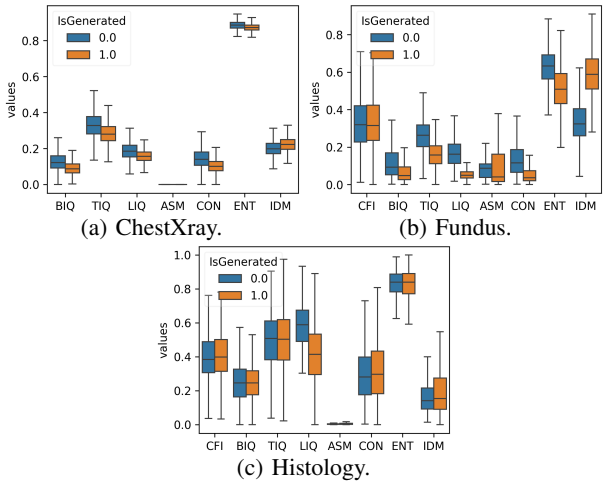
clarity (e.g., sharpness) of the input image; clearer images tend to have larger gradient values. In addition, we consider classic image texture features [15], which can be extracted from gray level concurrence matrix (GLCM). For this preliminary study, we describe image texture with 4 independent features in [15], namely, angular second moment (ASM), contrast (CON), entropy (ENT), and inverse different moment (IDM).

**Manifold Learning**. From a theoretical point of view, it is important to evaluate the quality and coverage of image generative models, in learning the manifolds of the training data. Recent research [16], [17] proposed to investigate the tradeoff between sample quality and variation by examining two measures, *precision* and *recall*, which provide more insights than Fréchet Inception Distance (FID) [18]. Intuitively, precision denotes the fraction of generated data that is realistic and recall denotes the fraction of the real data manifold covered by the generative model. The state-of-the-art approach [17], i.e., improved precision and recall (IPR), estimates the manifolds of real and generated data in a feature space induced by pre-trained deep neural networks (e.g., VGG-16 and Inception-V3). Practically, a volume is defined for each real/generated sample using its $k$-nearest neighbors in the real/generated set. The IPR metric examines how likely a generated/real sample is inside the volume of any real/generated sample, respectively. We set $k = 3$ as in [17] and adopted the features of the deepest layer of the Inception-V3 model to calculate the IPR metric.

## III. Results

**Improved Precision and Recall**. We first present the IPR metric [17] in Table I. Higher precision indicates that generated samples are more realistic, and higher recall indicates better coverage of the real manifold. The results suggest that the generative model for ChestXray may have better quality and coverage than models in other datasets. Generated Histology samples are less realistic (low precision) but may capture partial data variation (0.28 recall). The generative model for Fundus exhibits poor performance, yielding low values for both precision and recall. Table I also includes the FID metric (lower value indicates higher quality) which shows consistent results. Note that the IPR metric depends on the provided real and synthetic samples and the parameter $k$ for manifold approximation. A comprehensive parameter study may be conducted for future work.

**Spatial Features**. We further examine the spatial features between real and generated samples in each dataset. Note that we excluded CFI for ChestXray as samples are grayscale images. Feature distributions are reported in Figure 2. As can be seen, generated ChestXray and Fundus samples by DDPM models are less sharp than real samples (lower values in BIQ/TIQ/LIQ), and exhibit lower contrast (CON), lower, complexity (ENT), and smoother local patterns (IDM). DDPM generated Fundus images show a large difference in feature values compared to real images. On the other hand, StyleGAN2 generated Histology samples show similar feature values to real images, except for Laplacian gradients (LIQ) which are lower than real samples.



(a) ChestXray. (b) Fundus.

(c) Histology.

Fig. 2: Spatial features for real and generated images.

TABLE II: Classification of real vs. generated images

| Dataset | Feature Set | F1 |
|---|---|---|
| ChestXray | {BIQ, TIQ, LIQ, ASM, CON, ENT, IDM} | 0.71 |
| Fundus | {CFI, BIQ, TIQ, LIQ, ASM, CON, ENT, IDM} | 0.99 |
| Histology | {CFI, BIQ, TIQ, LIQ, ASM, CON, ENT, IDM} | 0.96 |

**Classification: real vs. generated**. We conducted a classification experiment by feeding the spatial features of real and generated images to a Support Vector Machine (SVM). The goal was to understand whether real data and generated data are separable in the feature space. We considered both linear and RBF kernels and a grid search was conducted to find the best parameters. 10% of data was used for training the classifiers and 90% was used for validation. F1 scores for three datasets are reported in Table II. It can be seen that spatial features can effectively separate real and generated images, especially for Fundus and Histology images. Although the F1 score for ChestXray is lower, we hypothesize that additional features, e.g., those extracted from color histograms [19], may further distinguish real and generated images.

## IV. Discussion and Conclusion

We presented a preliminary study on the performance of state-of-the-art generative models used in medical imaging. Empirical results show that existing generative models do not fully learn the training data manifolds in medical imaging, and generated samples differ very much from real samples in spatial features. While StyleGAN and DDPM models show superior performance to traditional generative models (e.g., Variational Auto-encoders and CGANs) in natural image domains, their applications in medical imaging domains, which come with higher data complexity and lower availability, have much room for improvement. This study also has its limitations. The quantitative measures reported may depend on the specific

samples used in computation (e.g., in manifold approximation with nearest neighbors for the IPR metric). Future work may consider varying both real and synthetic samples to obtain calibrated measures. Furthermore, future work may evaluate generative models toward target applications, e.g., glaucoma screening with fundus images.

## ACKNOWLEDGMENT

LF is supported in part by the National Science Foundation under grants CNS-2027114 and CNS-2144684. AB is primarily supported by the National Science Foundation under an REU supplement to CNS-2027114. LB is supported in part by a National Human Genome Research Institute grant R00HG010493, and a National Library of Medicine grant R01LM013712. The opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the sponsors.

## REFERENCES

[1] H. Wen, Y. Wu, J. Li, and H. Duan, "Communication-efficient federated data augmentation on non-iid data," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 3377–3386.

[2] F. Garcea, A. Serra, F. Lamberti, and L. Morra, "Data augmentation for medical imaging: A systematic literature review," *Computers in Biology and Medicine*, vol. 152, p. 106391, 2023.

[3] M. Pennisi, F. P. Salanitri, G. Bellitto, B. Casella, M. Aldinucci, S. Palazzo, and C. Spampinato, "Feder: Federated learning through experience replay and privacy-preserving data synthesis," *Computer Vision and Image Understanding*, vol. 238, p. 103882, 2024.

[4] M. J. Willemink, W. A. Koszek, C. Hardell, J. Wu, D. Fleischmann, H. Harvey, L. R. Folio, R. M. Summers, D. L. Rubin, and M. P. Lungren, "Preparing medical imaging data for machine learning," *Radiology*, vol. 295, no. 1, pp. 4–15, 2020.

[5] X. Yi, E. Walia, and P. Babyn, "Generative adversarial network in medical imaging: A review," *Medical image analysis*, vol. 58, p. 101552, 2019.

[6] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in neural information processing systems*, vol. 33, pp. 6840–6851, 2020.

[7] G. Müller-Franzes, J. M. Niehues, F. Khader, S. T. Arasteh, C. Haarburger, C. Kuhl, T. Wang, T. Han, T. Nolte, S. Nebelung *et al.*, "A multimodal comparison of latent denoising diffusion probabilistic models and generative adversarial networks for medical image synthesis," *Scientific Reports*, vol. 13, no. 1, p. 12098, 2023.

[8] R. Osuala, K. Kushibar, L. Garrucho, A. Linardos, Z. Szafranowska, S. Klein, B. Glocker, O. Diaz, and K. Lekadir, "A review of generative adversarial networks in cancer imaging: New applications, new solutions," *arXiv preprint arXiv:2107.09543*, pp. 1–64, 2021.

[9] J. Irvin, P. Rajpurkar, M. Ko, Y. Yu, S. Ciurea-Ilcus, C. Chute, H. Marklund, B. Haghgoo, R. Ball, K. Shpanskaya *et al.*, "Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 33, no. 01, 2019, pp. 590–597.

[10] C. De Vente, K. A. Vermeer, N. Jaccard, H. Wang, H. Sun, F. Khader, D. Truhn, T. Aimyshev, Y. Zhanibekuly, T.-D. Le *et al.*, "Airogs: Artificial intelligence for robust glaucoma screening challenge," *IEEE transactions on medical imaging*, 2023.

[11] A. Aksac, D. J. Demetrick, T. Ozyer, and R. Alhajj, "Brecahad: a dataset for breast cancer histopathological annotation and diagnosis," *BMC research notes*, vol. 12, no. 1, pp. 1–3, 2019.

[12] "Stylegan2 pretrained models," https://github.com/NVlabs/stylegan3, accessed: 2024-01-31.

[13] D. Hasler and S. E. Suesstrunk, "Measuring colorfulness in natural images," in *Human vision and electronic imaging VIII*, vol. 5007. SPIE, 2003, pp. 87–95.

[14] A. Mittal, A. K. Moorthy, and A. C. Bovik, "No-reference image quality assessment in the spatial domain," *IEEE Transactions on image processing*, vol. 21, no. 12, pp. 4695–4708, 2012.

[15] R. M. Haralick, K. Shanmugam, and I. H. Dinstein, "Textural features for image classification," *IEEE Transactions on systems, man, and cybernetics*, no. 6, pp. 610–621, 1973.

[16] M. S. Sajjadi, O. Bachem, M. Lucic, O. Bousquet, and S. Gelly, "Assessing generative models via precision and recall," *Advances in neural information processing systems*, vol. 31, 2018.

[17] T. Kynkäänniemi, T. Karras, S. Laine, J. Lehtinen, and T. Aila, "Improved precision and recall metric for assessing generative models," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[18] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," *Advances in neural information processing systems*, vol. 30, 2017.

[19] M. A. Stricker and M. Orengo, "Similarity of color images," in *Storage and retrieval for image and video databases III*, vol. 2420. SPiE, 1995, pp. 381–392.