Disentangling dataset size from synthetic diversity in tuberculosis chest X-ray classification

Connor Pink*, Edward R. Sykes
School of Computer Science, University of Guelph, Guelph, Ontario, Canada

Abstract

Synthetic data augmentation is often proposed as a remedy for limited and imbalanced medical imaging datasets. We study tuberculosis detection on the Tuberculosis Chest X-ray Database by training a 256×256 WGAN-GP and a 512×512 latent diffusion model fine-tuned from RoentGen-v2. We evaluated both for image quality and downstream utility. On generative metrics, diffusion outperforms WGAN-GP, achieving lower FID (6.56 vs. 9.28) and substantially lower radiology-aligned Rad-Dino FID (117.89 vs. 201.78), along with higher SSIM/MS-SSIM under our deterministic gen–real pairing protocol. However, in controlled DenseNet-121 classifier experiments under a fixed optimization budget (4,000 steps with identical selection criteria), synthetic augmentation does *not* outperform a count-matched duplicate-real control at matched dataset size. The duplicate-real control yields the best downstream performance despite adding no new information (e.g., 0.9981 ± 0.0013 test AUPRC at $r = 5$), while the best synthetic setting is diffusion at low ratio ($r = 0.25$). Increasing the synthetic-to-real ratio is not beneficial: high synthetic proportions degrade downstream performance, with particularly sharp deterioration for WGAN-GP at large ratios. Overall, the study demonstrates that superiority on generative metrics does not guarantee downstream benefit and highlights the importance of rigorous, count-matched augmentation controls when claiming gains from synthetic data.

Keywords: Generative AI, Diffusion Models, Tuberculosis, Medical Imaging, Data Augmentation, GAN

1. Introduction

Deep learning has improved medical-image analysis, but performance remains constrained by dataset size, class imbalance, and restricted data sharing [1–3]. Tuberculosis (TB) chest X-ray screening is a representative case: the public dataset used here contains roughly 3,500 normal radiographs and 700 TB-positive cases, making it useful but still limited for high-capacity models [4]. Generative models are often proposed as a remedy because they can synthesize plausible radiographs and enlarge the effective training distribution.

Prior work has shown that synthetic augmentation can help some medical-image classifiers and that strong Convolutional Neural Networks (CNN) backbones such as DenseNet perform well on thoracic imaging tasks [4–7]. Our question is narrower and methodological: *Does synthetic augmentation still help once effective sample count is controlled?* To answer this, we compare conditional WGAN-GP and latent diffusion (fine-tuned from RoentGen-v2 [8]) under matched downstream conditions on the TB Chest X-ray Database [4].

We frame the paper as a controlled evaluation study rather than an algorithm paper. Our contributions are: (i) a benchmark that matches downstream optimization conditions across generators and augmentation ratios; (ii) a count-matched duplicate-real control that separates synthetic-diversity effects from count and optimization effects; and (iii) evidence that better generative metrics, including radiology-aligned Rad-Dino FID, do not necessarily yield better downstream diagnostic utility. Across these experiments, synthetic augmentation does not outperform the duplicate-real control under a fixed optimization budget,

* corresponding author: pinkc@uoguelph.ca

indicating that apparent gains over real-only training are largely explained by reweighting and optimization effects rather than synthetic diversity.

2. Related Work

Generative modelling is widely used in medical imaging to address small and imbalanced datasets, with Generative Adversarial Networks (GANs) and diffusion models as the dominant approaches [3, 5, 9–11]. In TB chest X-ray analysis, prior work has shown both that established CNN backbones can achieve strong screening performance and that GAN- or diffusion-generated images can be used in augmentation pipelines [4, 6, 12]. Diffusion models are often reported to yield better fidelity than GANs in medical imaging [11, 13], but prior TB studies mostly ask whether synthetic images can help at all. We instead ask whether they help once count is matched with a duplicate-real control. For generative evaluation, we report conventional Fréchet Inception Distance (FID) and radiology-aligned RAD-DINO FID [14, 15]; the latter uses a more domain-aligned feature extractor but still remains a global embedding-distribution metric rather than a lesion-level clinical measure.

3. Methodology

This study evaluates whether synthetic TB chest X-rays improve downstream classification once count and optimization effects are controlled.

3.1. Dataset and Preprocessing

We utilized the Tuberculosis Chest X-ray Database [4], resized all images to 512×512, normalized intensities to $[0, 1]$, and used a reproducible 68/17/15 train/validation/test split. The test set was held out before all model development. Classical augmentation used horizontal flips, small rotations, and minor brightness shifts.

3.2. Generative Models

We trained a conditional WGAN-GP following Nascimento et al. [6] and a latent diffusion model obtained by fine-tuning the U-Net of RoentGen-v2 [8, 16]. WGAN-GP was trained at 256×256 and upsampled to 512×512 for downstream use, whereas diffusion operated directly at 512×512. Additional optimizer and training details are provided in the appendix.

For generative evaluation, we report FID, Rad-Dino FID, Structural Similarity Index (SSIM), MS-SSIM, and Learned Perceptual Image Patch Similarity (LPIPS). SSIM and MS-SSIM were computed under a deterministic within-class gen–real pairing protocol, while LPIPS was used as an intra-class diversity statistic over generated image pairs.

3.3. Augmented Classifier Training Experiments

To evaluate downstream utility, we trained DenseNet-121 classifiers [7], a standard chest X-ray backbone chosen to keep the downstream architecture fixed across all conditions. We compared real-only training, classical augmentation, a count-matched duplicate-real control, and synthetic augmentation with diffusion or WGAN-GP at ratios $r \in \{0.25, 0.5, 1, 2, 5\}$. Validation and test sets always contained only real images. All classifier runs used the same optimization budget (4,000 steps), selection criterion (validation AUPRC), and three random seeds so that differences reflect augmentation regime rather than unequal training conditions.

4. Results and Discussion

4.1. Generative Model Evaluation

Diffusion outperforms WGAN-GP on both FID (6.56 vs. 9.28) and radiology-aligned Rad-Dino FID (117.89 vs. 201.78), while also achieving higher SSIM/MS-SSIM under the same pairing protocol. Qualitatively, diffusion samples appear smoother and more structurally coherent, whereas WGAN-GP outputs are blurrier and less diverse.

4.2. Diagnostic Evaluation Using DenseNet-121

Training family	r	Val AUPRC	Test AUPRC	Test Sens.	Test F1(TB)	Test Bal. Acc.
Real only	0	0.9717 ± 0.0076	0.9829 ± 0.0068	0.9111 ± 0.0198	0.9348 ± 0.0065	0.9517 ± 0.0090
Real + classical aug.	0	0.9264 ± 0.0143	0.9477 ± 0.0115	0.8413 ± 0.0291	0.8745 ± 0.0212	0.9124 ± 0.0153
Real + duplicated (control)	5	0.9875 ± 0.0018	0.9981 ± 0.0013	0.9683 ± 0.0306	0.9696 ± 0.0117	0.9813 ± 0.0144
Real + diffusion synthetic	0.25	0.9785 ± 0.0098	0.9900 ± 0.0098	0.9333 ± 0.0252	0.9407 ± 0.0156	0.9616 ± 0.0126
Real + WGAN-GP synthetic	0.25	0.9697 ± 0.0132	0.9822 ± 0.0112	0.9079 ± 0.0334	0.9240 ± 0.0103	0.9483 ± 0.0141

Table 1. Downstream DenseNet-121 performance (mean ± std across 3 seeds) for five training families. r denotes the synthetic-to-real (or duplicate-to-real) ratio selected by mean validation AUPRC within each family. Validation and test sets contain only real images.

Configuration	r	Test AUPRC	95% CI	Test F1(TB)	95% CI
Real only (dev baseline)	0	0.9829 ± 0.0068	–	0.9348 ± 0.0065	–
Real + duplicated (final)	5	0.9992 ± 0.0005	$[0.9971, 1.0000]$	0.9444 ± 0.0139	$[0.9117, 0.9731]$
Real + diffusion synthetic (final)	0.25	0.9982 ± 0.0009	$[0.9949, 1.0000]$	0.9587 ± 0.0121	$[0.9275, 0.9830]$

Table 2. Final retraining results with bootstrap 95% confidence intervals on the test set for the best overall and best synthetic configurations. Real-only is shown for reference. Metrics are reported as mean ± std across 3 seeds.

The downstream experiments compare real-only training, classical augmentation, synthetic augmentation, and a count-matched duplicate-real control designed to isolate count and reweighting effects from any benefit due to synthetic diversity. The main result is that the **real-only baseline is already near ceiling** (Table 1), the **duplicate-real control performs best despite adding no new information**, and **diffusion outperforms WGAN-GP but still does not surpass duplication**. Table 1 gives the full five-family summary: the real-only baseline achieves 0.9829 ± 0.0068 test AUPRC, the duplicate-real control reaches 0.9981 ± 0.0013 at $r = 5$, and diffusion is the strongest synthetic method at 0.9900 ± 0.0098 with $r = 0.25$.

The scaling trend is also clear: **more synthetic data is not monotonically better**. Diffusion peaks at a low synthetic proportion and then degrades as synthetic samples dominate training, while WGAN-GP degrades sharply at larger ratios. This pattern suggests that gains over real-only are driven more by **optimization/reweighting effects** than by added synthetic diversity.

Figure 1 gives the cleanest test of this claim by comparing synthetic augmentation directly against the duplicate-real control at matched effective dataset size.

Across all ratios, both generators underperform the duplicate-real control in test AUPRC at matched effective dataset size (Figure 1). For diffusion, the gap is small at low ratio but grows as synthetic images dominate training; for WGAN-GP, the gap becomes extreme at larger ratios, consistent with downstream collapse when synthetic samples substantially outnumber real samples. This is the clearest evidence that observed

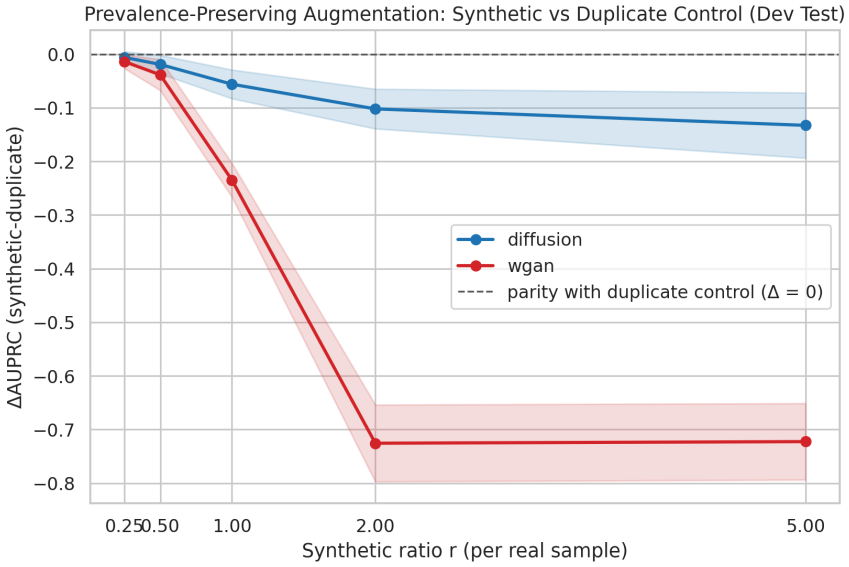

Figure 1. 'ΔAUPRC' (synthetic-duplicate) vs ratio; dashed 'Δ=0' line marks parity with duplicate control.

gains over real-only do not, by themselves, justify claims of benefit from *synthetic diversity*. Final confirmatory confidence intervals are reported in Table 2.

Because WHO TB screening guidance emphasizes high sensitivity [17], it is also useful to inspect how TB recall changes with the augmentation ratio. Figure 2 shows that the real-only baseline already exceeds 90% sensitivity on average, classical augmentation falls below that reference level, and the duplicate-real control and low-ratio diffusion setting remain above it across seeds. As with the other downstream metrics, performance worsens as synthetic images make up a larger share of the training set, especially for WGAN-GP.

Overall, diffusion produces higher-fidelity synthetic CXRs than WGAN-GP and yields the strongest synthetic augmentation results at low ratio, but synthetic augmentation still does not outperform a count-matched duplicate-real control. Because the held-out split is already easy for a strong baseline, augmentation headroom is small, so these null synthetic gains should not be overgeneralized to much harder or lower-data settings. At the same time, this is exactly where strict controls matter most: when baseline performance is already high, apparent augmentation gains can easily reflect optimization or sample-count effects rather than new information. Limitations remain. The split is near ceiling; all comparisons are under a fixed optimization budget; minority-class-only augmentation is not isolated with duplicate-TB controls; reported generative metrics remain global rather than lesion-aware; and WGAN-GP is additionally disadvantaged by a 256×256 to 512×512 resolution mismatch.

5. Conclusion

We conducted a controlled comparison of conditional WGAN-GP and latent diffusion (RoentGen-v2 fine-tuning) for TB chest X-ray synthesis. Diffusion outperforms WGAN-GP on generative metrics and yields the stronger synthetic augmentation results, but the central downstream finding is methodological: **under a fixed optimization budget, synthetic augmentation does not outperform a count-matched duplicate-real**

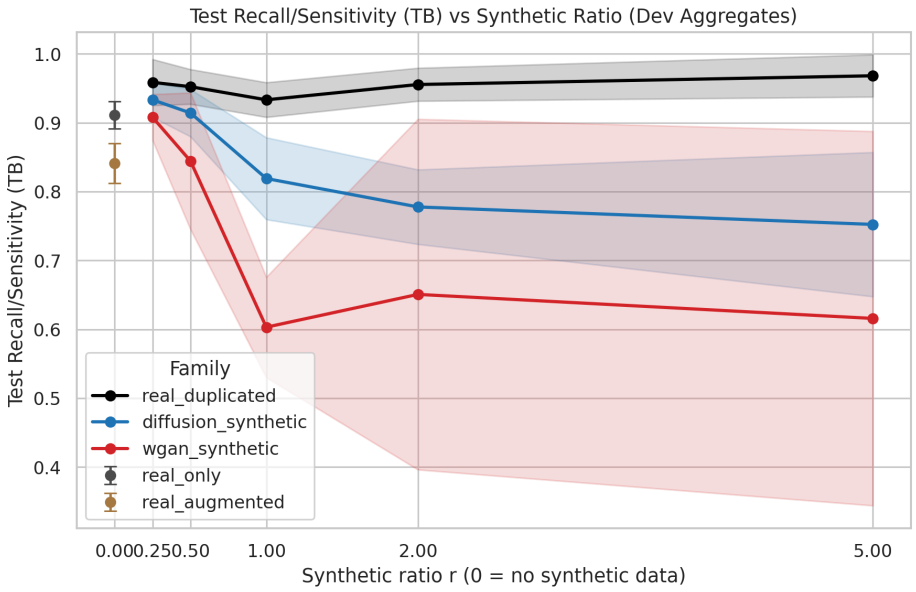

Figure 2. Test sensitivity (TB recall) vs. ratio r (mean $\pm$ std across seeds). Sensitivity remains strongest for the duplicate-real control and low-ratio diffusion, while higher synthetic proportions degrade recall, especially for WGAN-GP.

control. The strongest gains over real-only training are achieved by duplicating real samples, implying that improvements are driven primarily by **reweighting and optimization effects** rather than synthetic diversity.

High synthetic-to-real proportions degrade performance, especially for WGAN-GP. More broadly, superior generative metrics do not guarantee downstream benefit, and rigorous count-matched controls are essential when evaluating synthetic augmentation claims. Future work should test targeted TB-positive augmentation, lower-data regimes, alternative downstream backbones, and broader cross-dataset validation.

Acknowledgements

This paper builds on an earlier course project. Thanks to Mirza Ahmadi and Yeganeh Jamshidi for their participation in that initial project, and Dr. Abdulrahman Al-Shanoon for helpful advice during the early stages of the work.

References

[1] X. Wang, H. Luo, and J. Gao. "Improving tuberculosis detection from chest X-rays with hybrid GAN-CNN approaches". In: *Computers in Biology and Medicine* 157 (2023), p. 106846. DOI: 10.1016/j.compbiomed.2023.106846.

[2] L. Wang, M. S. Khosravi, and D. Rueckert. "Generative Artificial Intelligence in Medical Imaging: Foundations, Progress, and Clinical Translation". In: *Nature Biomedical Engineering* 9 (2025), pp. 1212–1234. DOI: 10.1038/s41551-025-01245-x.

[3] C. Yi, Q. Zhou, J. Ding, and Y. Zhao. "Diffusion-based data augmentation for medical image classification under limited datasets". In: *Medical Image Analysis* 91 (2025), p. 103915. DOI: 10.1016/j.media.2025.103915.

[4] T. Rahman, A. Khandakar, M. A. Kadir, K. R. Islam, K. F. Islam, R. Mazhar, T. Hamid, M. T. Islam, Z. B. Mahbub, M. A. Ayari, and M. E. H. Chowdhury. "Reliable Tuberculosis Detection using Chest X-ray with Deep Learning, Segmentation and Visualization". In: *IEEE Access* 8 (2020), pp. 191586–191601. DOI: 10.1109/ACCESS.2020.3031384. URL: https://doi.org/10.1109/ACCESS.2020.3031384.

[5] L. Frid-Adar, E. Klang, M. Amitai, J. Goldberger, and H. Greenspan. "Synthetic data augmentation using GAN for improved liver lesion classification". In: *IEEE Transactions on Medical Imaging* 38.3 (2018), pp. 731–739. DOI: 10.1109/TMI.2018.2867350.

[6] O. T. Nascimento, J. M. de Seixas, and A. Trajman. "Synthetic chest X-ray data generation for tuberculosis infection detection using generative adversarial networks". In: *Neural Computing and Applications* 37 (2025), pp. 18151–18171. DOI: 10.1007/s00521-024-09811-9.

[7] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger. "Densely Connected Convolutional Networks". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017.

[8] S. L. Moroianu, C. Bluethgen, P. Chambon, M. Cherti, J.-B. Delbrouck, M. Paschali, B. Price, J. Gichoya, J. Jitsev, C. P. Langlotz, and A. S. Chaudhari. "Improving Performance, Robustness, and Fairness of Radiographic AI Models with Finely-Controllable Synthetic Data". In: *arXiv preprint arXiv:2508.16783* (2025). URL: https://arxiv.org/abs/2508.16783.

[9] T. Karras, S. Laine, and T. Aila. "A style-based generator architecture for generative adversarial networks". In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2019), pp. 4401–4410. DOI: 10.1109/CVPR.2019.00453.

[10] J. Ho, A. Jain, and P. Abbeel. "Denoising diffusion probabilistic models". In: *Advances in Neural Information Processing Systems (NeurIPS)*. Vol. 33. 2020, pp. 6840–6851.

[11] A. Kazerouni, M. van der Hout, and M. Tschannen. "Diffusion models for medical imaging: a review". In: *IEEE Access* 11 (2023), pp. 97812–97834. DOI: 10.1109/ACCESS.2023.3294742.

[12] D. I. Morís, J. de Moura, J. Novo, and M. Ortega. "Adapted Generative Latent Diffusion Models for Accurate Pathological Analysis in Chest X-ray Images". In: *Medical & Biological Engineering & Computing* 62.7 (July 1, 2024), pp. 2189–2212. ISSN: 1741-0444. DOI: 10.1007/s11517-024-03056-5. URL: https://doi.org/10.1007/s11517-024-03056-5 (visited on 02/05/2026).

[13] P. Dhariwal and A. Nichol. "Diffusion models beat GANs on image synthesis". In: *Advances in Neural Information Processing Systems (NeurIPS)* 34 (2021), pp. 8780–8794.

[14] D. A. Chan and S. P. Sithungu. "Evaluating the Suitability of Inception Score and Fréchet Inception Distance as Metrics for Quality and Diversity in Image Generation". In: *Proceedings of the 2024 7th International Conference on Computational Intelligence and Intelligent Systems*. CIIS '24. New York, NY, USA: Association for Computing Machinery, 2025, 79–85. ISBN: 9798400717437. DOI: 10.1145/3708778.3708790. URL: https://doi.org/10.1145/3708778.3708790.

[15] Microsoft. *rad-dino (Revision 72881b8)*. 2024. DOI: 10.57967/hf/3050. URL: https://huggingface.co/microsoft/rad-dino.

[16] P. Chambon, C. Bluethgen, C. P. Langlotz, and A. Chaudhari. *Adapting Pretrained Vision-Language Foundational Models to Medical Imaging Domains*. Oct. 9, 2022. DOI: 10.48550/arXiv.2210.04133. arXiv: 2210.04133 [cs]. URL: http://arxiv.org/abs/2210.04133 (visited on 02/12/2026). Pre-published.

[17] W. H. Organization. *High priority target product profiles for new tuberculosis diagnostics: report of a consensus meeting, 28–29 April 2014, Geneva, Switzerland*. WHO Technical Report WHO/HTM/TB/2014.18. Licence: CC BY-NC-SA 3.0 IGO. Accessed 2024-12-10. Geneva, Switzerland: World Health Organization, 2014. URL: https://www.who.int/publications/i/item/WHO-HTMTB-2014.18.

Appendix A. Training Details

The conditional WGAN-GP used a learned class embedding, 256×256 training resolution, batch size 32, and gradient penalty weight $\lambda_{gp} = 10$. The latent diffusion model was fine-tuned from RoentGen-v2 at 512×512 with frozen VAE and text encoder, batch size 4, and a maximum of 15,000 optimization steps. In downstream classifier experiments, all conditions used DenseNet-121, batch size 8, learning rate 1×10^{-4}, validation every 100 steps, early stopping based on validation AUPRC, and three random seeds.

Appendix B. Generative Samples

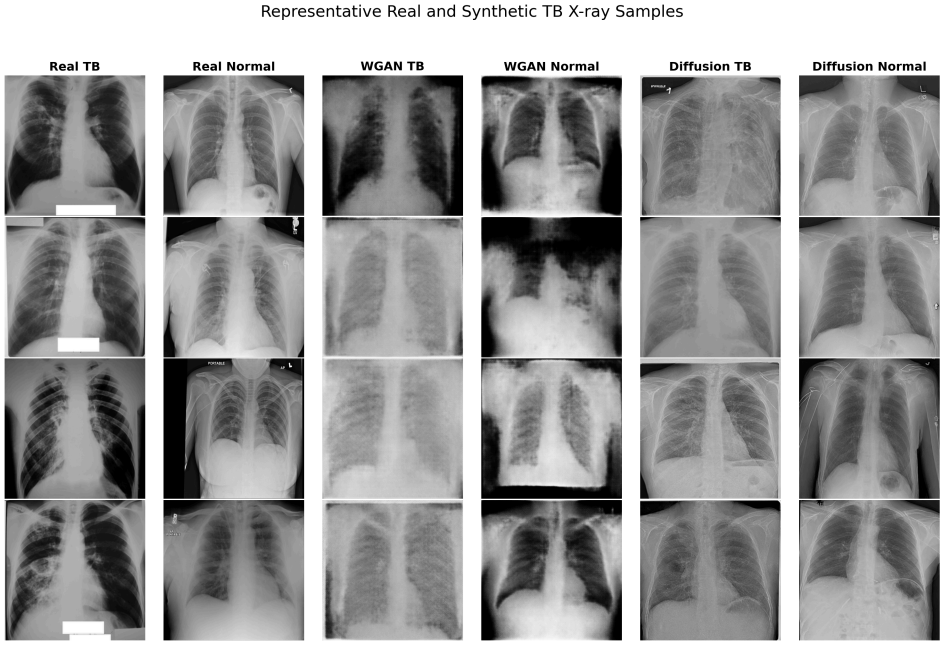

Figure 3. TB and normal sample images from real, WGAN-generated, and diffusion-generated datasets.

Appendix C. Supplementary Downstream Results

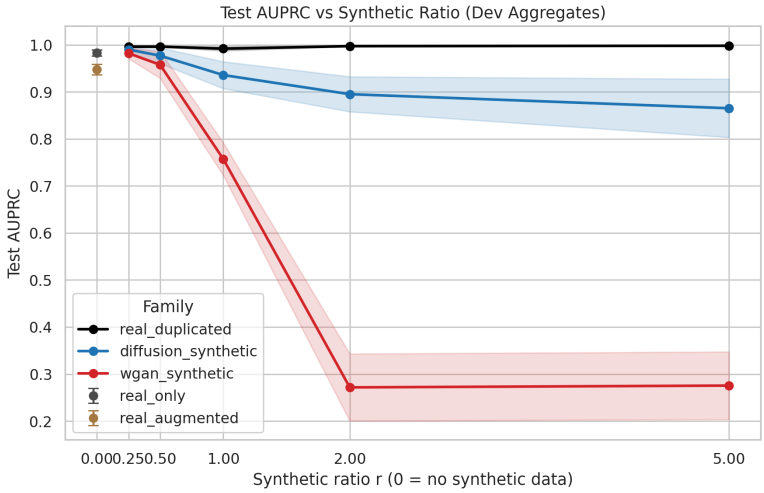

Figure 4. Test AUPRC vs. synthetic-to-real ratio r (mean $\pm$ std across seeds) for diffusion and WGAN-GP augmentation, compared to the count-matched duplicate-real control. Increasing synthetic proportion does not yield monotonic improvements; diffusion performs best at low r, while WGAN-GP degrades substantially at higher r.