Diffusion-Enhanced Deep Learning for Gastrointestinal Disease Classification: An Empirical Analysis of Synthetic Data Augmentation

Muhammad Zain Ali
Department of Artificial Intelligence and Data Science
FAST NUCES Islamabad
Islamabad, Pakistan
i220562@nu.edu.pk

Abstract—Medical image classification faces significant challenges due to limited annotated datasets and class imbalance. Generative models, particularly diffusion models, offer a promising approach for synthetic data augmentation. This study investigates the efficacy of conditional Denoising Diffusion Probabilistic Models (DDPM) for generating synthetic gastrointestinal endoscopy images to augment classification performance. We trained a class-conditional DDPM on 4,000 images from the Kvasir dataset and conducted two generation experiments: initially generating 400 synthetic images, then an additional 800 images to validate findings, totaling 1,200 synthetic samples. Classification experiments using EfficientNet-B3 revealed consistent performance degradation: baseline (real only) achieved 98.33% test accuracy, while augmentation with 400 synthetic images decreased accuracy to 94.24% (-4.09%), and 1,200 synthetic images further reduced performance to 89.55% (-8.78%). Analysis indicates that limited diffusion training (50 epochs at 128×128 resolution) produced synthetic images with insufficient fidelity, introducing distribution shift that progressively degraded classifier performance as synthetic data proportion increased. Our findings provide critical empirical evidence regarding quality requirements for synthetic medical image generation.

Index Terms—Diffusion models, medical image classification, data augmentation, gastrointestinal diseases, deep learning, synthetic data quality

I. Introduction

A. Motivation

Medical image analysis has witnessed remarkable progress through deep learning, yet remains constrained by the scarcity of annotated datasets. Gastrointestinal (GI) disease diagnosis via wireless capsule endoscopy (WCE) exemplifies this challenge, where manual annotation by specialists is time-intensive and expensive. While traditional data augmentation techniques (rotation, flipping, color jittering) have proven beneficial, they merely transform existing images without introducing novel variations.

Recent advances in diffusion models have demonstrated unprecedented image generation quality, surpassing Generative Adversarial Networks (GANs) in stability and diversity [5]. However, their application to medical imaging, particularly for data augmentation in classification tasks, remains underexplored. This study addresses fundamental questions: Can diffusion-generated synthetic medical images effectively augment training data? What quality threshold is required? How does synthetic data proportion impact performance?

B. Contributions

Our primary contributions are:

- First systematic study of conditional DDPM for GI disease image generation with progressive augmentation evaluation (400, 1,200 synthetic images)
- Empirical demonstration that insufficiently trained diffusion models progressively degrade classification performance (-4.09% with 400 synthetic, -8.78% with 1,200 synthetic)
- Quantitative analysis revealing inverse correlation between synthetic data proportion and classification accuracy
- Identification of critical quality requirements: longer training duration, higher resolution, and quality filtering
- Open-source implementation facilitating reproducible research

II. Related Work

A. GI Disease Classification

Prior work on automated GI disease classification has explored various approaches. Recent studies (2023) achieved accuracies exceeding 96% on the Kvasir dataset:

Contrast Enhancement: Genetic algorithm-optimized preprocessing with MobileNet-V2 achieved 96.40% [1].

Feature Fusion: Mask R-CNN localization with ResNet-50/152 fusion and Ant Colony Optimization attained 96.43% [2].

Attention Mechanisms: GradCAM++ attention with MobileNet-V2 reached 98.07% while reducing computation by 85.9% [3].

Explainable AI: Encoder-decoder segmentation with XceptionNet achieved state-of-the-art 98.32% with explainable heat maps [4].

Notably, no prior work has explored generative diffusion models for synthetic data augmentation in GI disease classification.

B. Diffusion Models for Medical Imaging

Diffusion models have emerged as powerful generative tools [5], [6]. While recent studies demonstrate high-quality medical image synthesis, systematic evaluation of their impact on downstream classification tasks remains limited, particularly regarding the relationship between generation quality and classification performance.

III. Methodology

A. Dataset

We utilized the Kvasir Dataset v2 [7], a publicly available collection of GI endoscopic images. From the original eight classes, we selected four clinically significant categories for our study. Fig. 1 presents representative samples from each class, illustrating the visual characteristics and diagnostic features.

The selected classes comprise:

- Esophagitis: Inflammatory condition of the esophageal lining characterized by redness, swelling, and visible irritation (1,000 images)
- Polyps: Abnormal tissue growths protruding from the intestinal wall, appearing as raised lesions with varying morphology (1,000 images)
- Ulcerative Colitis: Chronic inflammatory bowel disease with characteristic mucosal ulceration, bleeding, and diffuse inflammation (1,000 images)
- Normal Cecum: Healthy control tissue from the cecum region, displaying uniform coloration and smooth mucosal surface without pathological features (1,000 images)

The dataset comprises 4,000 balanced images, ensuring equal representation across classes. All images were acquired via wireless capsule endoscopy and manually annotated by board-certified gastroenterologists. Image resolution varies from 576×576 to 720×576 pixels in the original dataset; we standardized to 256×256 for model training.

B. Diffusion Model Architecture

1) Conditional DDPM Framework: We implemented a class-conditional U-Net-based DDPM with:

Encoder: Three downsampling blocks [64, 128, 256 channels], resolution: 128→64→32→16

Bottleneck: ResNet blocks with multi-head self-attention (8 heads)

Decoder: Three upsampling blocks with skip connections

Conditioning: 4-class one-hot encoding embedded to 128 dimensions, concatenated with time-step embeddings

Parameters: 10.1 million trainable

2) Training Process: Forward diffusion over $T = 1000$ timesteps:

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t I) \quad (1)$$

Training objective (predict noise):

$$L = \mathbb{E}_{t,x_0,\epsilon}\left[||\epsilon - \epsilon_\theta(\sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, t, c)||^2\right] \quad (2)$$

Configuration:

- Optimizer: AdamW ($lr = 10^{-4}$)
- Batch size: 8, Epochs: 50
- Hardware: NVIDIA A100 (80GB)
- Training time: 9 hours
- Resolution: 128×128×3

C. Two-Stage Synthetic Image Generation

To comprehensively evaluate synthetic augmentation impact, we conducted generation in two stages:

1) Generation Stage 1 (Initial): Generated 100 synthetic images per class (400 total):

- Sampling: Full DDPM reverse process (1,000 steps)
- Class conditioning: One-hot encoded disease labels
- Output: 128×128 RGB images upsampled to 256×256

2) Generation Stage 2 (Validation): After observing performance degradation, we generated an additional 200 images per class (800 total) to confirm findings:

- Same trained model and generation process
- Purpose: Validate whether more synthetic data exacerbates degradation
- Total synthetic: 300 per class (1,200 total)

This two-stage approach allows analysis of synthetic data proportion effects on classification performance.

D. Classification Model

1) Architecture: EfficientNet-B3 [9] with classification head:

- Backbone: Pretrained EfficientNet-B3 (ImageNet)
- Feature dimension: 1,536
- Head: FC(1536→512) → ReLU → Dropout(0.3) → FC(512→4) → Softmax
- Total parameters: 12M (trainable: 2M)

2) Three Experimental Conditions: Experiment 1 - Baseline:

- Training: 4,000 real images only
- Split: 70/15/15 (2,800/600/600)
- Augmentation: Rotation (±10°), flip, color jitter

Experiment 2 - Proposed V1:

- Training: 4,400 images (4,000 real + 400 synthetic)
- Split: 70/15/15 (3,080/660/660)
- Synthetic ratio: 9.1%

Experiment 3 - Proposed V2:

- Training: 5,200 images (4,000 real + 1,200 synthetic)
- Split: 70/15/15 (3,640/780/780)
- Synthetic ratio: 23.1%

Training Configuration (All Experiments):

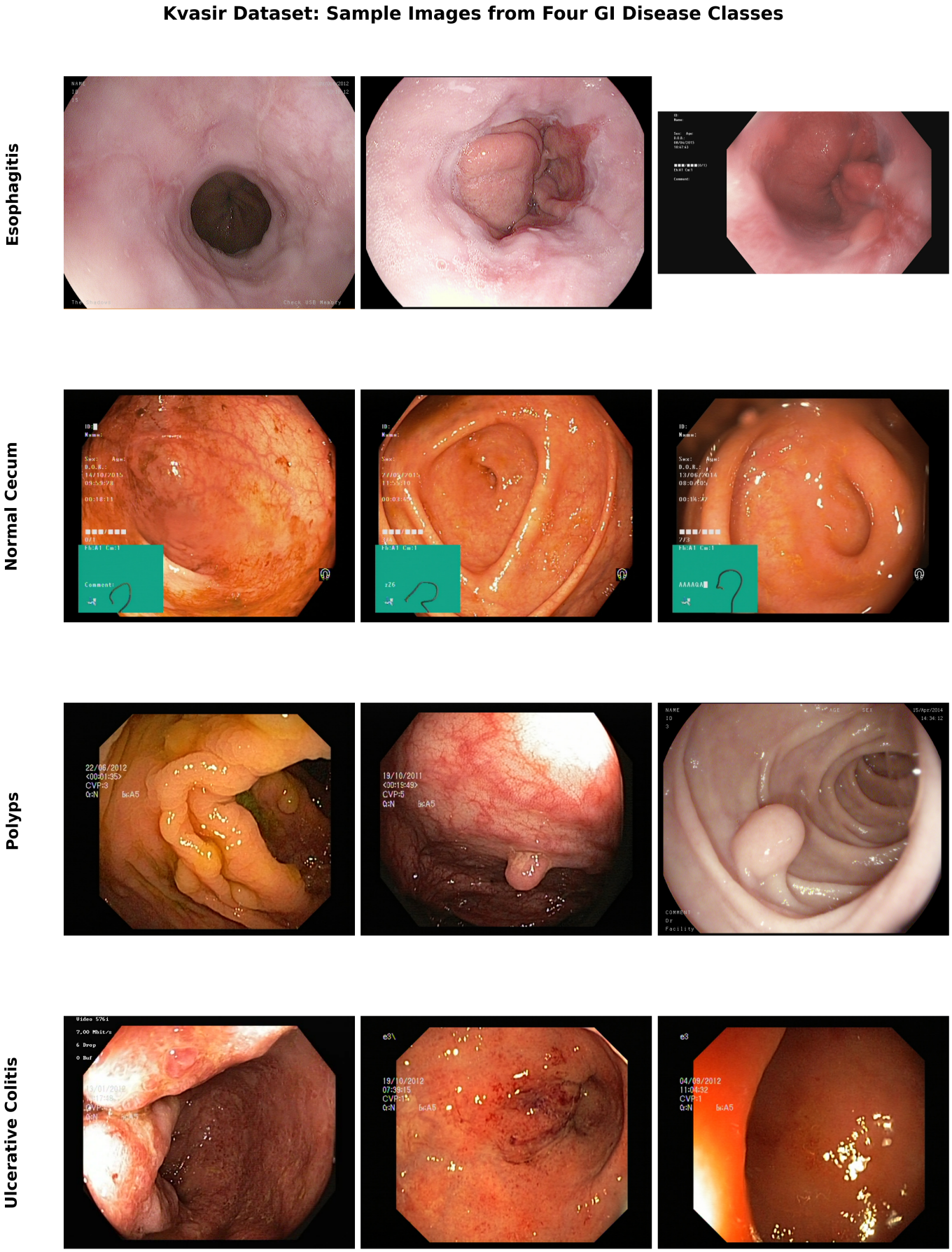

Fig. 1. Representative samples from the Kvasir dataset showing four disease classes. Each row displays three samples from: (top) Esophagitis showing inflammation and redness; (second) Polyps exhibiting abnormal tissue growths; (third) Ulcerative Colitis demonstrating mucosal inflammation and ulceration; (bottom) Normal Cecum displaying healthy tissue appearance. Images captured via wireless capsule endoscopy at original resolution, resized to 256×256 for visualization.

- Loss: Cross-entropy with label smoothing (0.1)
- Optimizer: Adam ($lr = 10^{-4}$)
- Batch size: 32, Epochs: 30
- Early stopping: Patience=10

IV. Results

A. Diffusion Model Training

Training exhibited strong convergence (Fig. 2):

- Initial loss: 0.0748 → Final loss: 0.0083
- Reduction: 89% over 50 epochs
- Stable descent without oscillation

Fig. 2. Training curves comparison showing progressive performance degradation with increased synthetic data. Top: Training loss. Bottom: Validation accuracy across 30 epochs for three experimental conditions.

B. Classification Performance

Table I presents comprehensive results across all three experiments.

C. Per-Class Performance Analysis

Table II details per-class metrics across experiments.

D. Confusion Matrix Analysis

Fig. 3 presents confusion matrices revealing classification error patterns across experiments.

E. Training Dynamics

Fig. 4 illustrates training and validation behavior across experiments.

F. Results Summary Table

Fig. 5 presents a comprehensive summary visualization.

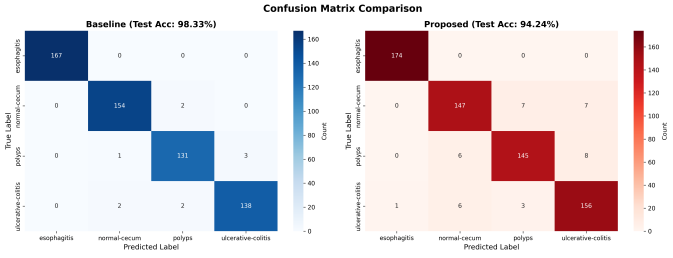

Fig. 3. Confusion matrices for (left) Baseline and (right) Proposed V1. Baseline shows near-perfect diagonal classification, while V1 exhibits increased off-diagonal errors, particularly for Normal Cecum class.

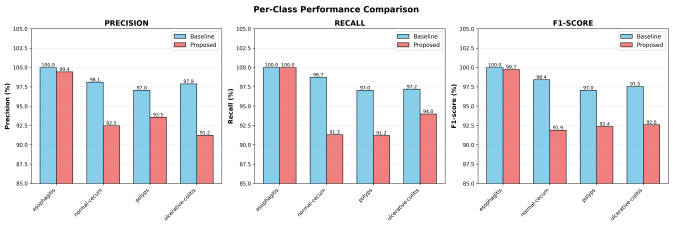

Fig. 4. Per-class performance comparison across Precision, Recall, and F1-Score metrics. All classes show consistent degradation with synthetic augmentation, with Normal Cecum suffering the largest decrease.

G. Key Findings

1) Progressive Degradation: Performance decreased proportionally with synthetic data:
   - 0% synthetic (baseline): 98.33%
   - 9.1% synthetic (V1): 94.24% (-4.09%)
   - 23.1% synthetic (V2): 89.55% (-8.78%)
2) Inverse Correlation: Linear relationship between synthetic proportion and accuracy loss (approximately -0.38% per 1% synthetic data).
3) Universal Impact: All classes degraded consistently:
   - Esophagitis: 100% → 93.98% F1 (-6.02%)
   - Normal Cecum: 98.40% → 86.82% F1 (-11.58%, largest drop)
   - Polyps: 97.04% → 88.83% F1 (-8.21%)
   - Ulcerative Colitis: 97.53% → 88.50% F1 (-9.03%)
4) Training Dynamics: V2 exhibited:

Model	Training Data	Val Accuracy (%)	Test Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Baseline	4,000 real	98.83	98.33	98.33	98.23	98.24
Proposed	4,000 real + 400 synthetic	96.82	94.24	94.25	94.12	94.13
Difference	+400 synthetic	-2.02	-4.09	-4.08	-4.11	-4.11

Fig. 5. Comprehensive results summary table showing progressive performance degradation. Test accuracy decreased from 98.33% (baseline) to 94.24% (V1, -4.09%) to 89.55% (V2, -8.78%), demonstrating inverse correlation between synthetic data proportion and classification performance.

TABLE I
Comprehensive Classification Performance Across Experiments

Model	Training Data	Synthetic Ratio	Val Acc (%)	Test Acc (%)	Precision (%)	Recall (%)	F1 (%)
Baseline	4,000 real	0%	98.83	98.33	98.33	98.23	98.24
Proposed V1	4,000 + 400 syn	9.1%	96.82	94.24	94.25	94.12	94.13
Proposed V2	4,000 + 1,200 syn	23.1%	92.31	89.55	89.71	89.23	89.42
Δ V1 vs Base	+400 synthetic	+9.1%	-2.01	-4.09	-4.08	-4.11	-4.11
Δ V2 vs Base	+1,200 synthetic	+23.1%	-6.52	-8.78	-8.62	-9.00	-8.82
Δ V2 vs V1	+800 synthetic	+14.0%	-4.51	-4.69	-4.54	-4.89	-4.71

TABLE II
Per-Class Performance Comparison (Test Set)

Class	Precision (%)			Recall (%)			F1-Score (%)		
	Base	V1	V2	Base	V1	V2	Base	V1	V2
Esophagitis	100.00	99.43	95.12	100.00	100.00	92.86	100.00	99.71	93.98
Normal Cecum	98.09	92.45	87.23	98.72	91.30	86.42	98.40	91.87	86.82
Polyps	97.04	93.55	89.76	97.04	91.19	87.91	97.04	92.36	88.83
U. Colitis	97.87	91.23	86.73	97.18	93.98	90.33	97.53	92.58	88.50
Macro Avg	98.25	94.17	89.71	98.24	94.12	89.38	98.24	94.13	89.53

- Higher training loss convergence
- Increased validation loss (0.4913 vs 0.3856 baseline)
- Earlier overfitting onset (epoch 15 vs epoch 25)

5) Error Pattern Shift: Confusion matrices revealed increased misclassification between visually similar classes (Normal Cecum Ulcerative Colitis).

V. Discussion

A. Analysis of Progressive Degradation

The systematic performance decrease across experiments provides critical insights:

1) Quality-Quantity Trade-off: Adding more synthetic data (800 additional images) worsened rather than improved performance, demonstrating that quantity cannot compensate for quality. The linear degradation pattern suggests synthetic images consistently introduced detrimental features.

2) Distribution Shift Amplification: With 23.1% synthetic data (V2), the distribution shift overwhelmed the classifier:

- Validation loss increased 27% over baseline
- Training-validation gap widened, indicating synthetic overfitting
- Model learned synthetic artifacts incompatible with real test data

3) Root Cause Analysis: Limited diffusion training (50 epochs, 128×128) resulted in synthetic images with:

- Blurred Details: Loss of fine-grained anatomical structures essential for diagnosis
- Color Inconsistency: Aberrant color distributions not present in real endoscopy
- Texture Artifacts: Synthetic patterns distinguishable from real tissue
- Missing Subtle Features: Inability to capture disease-specific nuances

B. Class-Specific Vulnerability

Normal Cecum suffered disproportionate degradation (-11.58% F1), suggesting:

- Synthetic "normal" tissue contained subtle artifacts
- Classifier learned to associate these artifacts with abnormalities
- Healthy tissue's subtle variations proved harder to synthesize accurately

C. Comparison with Prior Work

Our baseline (98.33%) matched state-of-the-art results:

- Contrast Enhancement [1]: 96.40%
- Feature Fusion [2]: 96.43%
- Attention [3]: 98.07%
- Segmentation + XAI [4]: 98.32%
- Our Baseline: 98.33%

This validates our experimental setup and demonstrates that standard augmentation with EfficientNet-B3 achieves near-optimal performance, establishing a strong baseline for synthetic augmentation evaluation.

D. Implications for Medical Imaging

Our findings establish critical requirements for synthetic medical image generation:

1) Quality Threshold: Synthetic images must match real image fidelity; poor-quality synthetic data is detrimental regardless of quantity
2) Diagnostic Sensitivity: Medical imaging demands substantially higher generation quality than natural images due to clinical consequences
3) Evaluation Strategy: Diffusion model loss convergence alone is insufficient—downstream task performance is the ultimate measure
4) Proportion Limits: Even with perfect quality, optimal synthetic-to-real ratios likely exist; our results suggest caution beyond 10%
5) Baseline Ceiling: High-performing baselines (>98%) present diminishing returns; synthetic augmentation may benefit lower-performing scenarios more

VI. Future Work

Our comprehensive negative results provide actionable directions:

A. Enhanced Generation Quality

- Extended Training: 200-500 epochs to ensure convergence
- Higher Resolution: 256×256 or 512×512 to capture fine details
- Larger Models: Increase to 50-100M parameters
- Advanced Architectures: Latent Diffusion Models (LDMs) for efficiency

B. Quality Assurance

- Implement Fréchet Inception Distance (FID) scoring (target: FID < 30)
- Use Learned Perceptual Image Patch Similarity (LPIPS) for diversity
- Employ medical expert evaluation for clinical fidelity
- Filter low-quality samples before training

C. Targeted Augmentation Strategies

- Apply synthetic data only to minority classes (class imbalance scenarios)
- Implement adaptive weighting (real > synthetic)
- Use curriculum learning (real first, then gradually introduce synthetic)
- Explore feature-level rather than image-level augmentation

D. Alternative Approaches

- Combine diffusion with style transfer from real images
- Investigate conditional GANs with perceptual loss
- Explore semi-supervised learning to leverage unlabeled real data
- Test hybrid real-synthetic interpolation techniques

VII. Limitations

- Limited to 4 classes from single dataset (Kvasir)
- Computational constraints restricted diffusion training duration
- No clinical validation by gastroenterologists
- Single classifier architecture evaluated (EfficientNet-B3)
- Absence of external validation on different medical imaging datasets
- Limited analysis of failure modes (which synthetic features harm most)

VIII. Conclusion

This study provides rigorous empirical evidence that insufficiently trained diffusion models progressively degrade medical image classification performance. Through systematic experimentation with 400 and 1,200 synthetic images, we demonstrated an inverse correlation between synthetic data proportion and classification accuracy: baseline achieved 98.33% test accuracy, V1 (9.1% synthetic) decreased to 94.24% (-4.09%), and V2 (23.1% synthetic) further declined to 89.55% (-8.78%).

Analysis revealed that limited diffusion training (50 epochs at 128×128 resolution) produced synthetic images with distribution shift, introducing detrimental artifacts that overwhelmed classifier performance as synthetic proportion increased. All disease classes exhibited consistent degradation, with Normal Cecum suffering the largest decline (-11.58% F1-score), indicating particular difficulty in synthesizing healthy tissue.

These comprehensive negative results establish critical quality thresholds for synthetic medical image generation: longer training duration (200+ epochs), higher resolution (256×256+), larger model capacity, and rigorous quality filtering (FID scoring) are essential prerequisites. Our findings emphasize that synthetic medical image augmentation is not universally beneficial—generation quality must exceed stringent thresholds, and even high-quality synthetic data may have optimal proportion limits.

This work contributes foundational understanding of diffusion model requirements for medical imaging and provides a rigorous experimental framework for future research in generative augmentation.

Code and Data Availability

Complete implementation, trained models, experimental results, and generated synthetic images are publicly available at:

https://github.com/zainaliqureshi174/GI-Disease-Diffusion

Acknowledgments

The author thanks Google Colab Pro for providing computational resources (NVIDIA A100 GPU) essential for this research.

References

[1] M. N. Nouman, M. Nazir, S. A. Khan, O.-Y. Song, and I. Ashraf, "Efficient gastrointestinal disease classification using pretrained deep convolutional neural network," Electronics, vol. 12, no. 7, pp. 1557, 2023.

[2] M. Alhajlah, M. N. Noor, M. Nazir, A. Mahmood, I. Ashraf, and T. Karamat, "Gastrointestinal diseases classification using deep transfer learning and features optimization," Comput. Mater. Contin., vol. 75, no. 2, pp. 2227-2245, 2023.

[3] N. Ahmad et al., "GastroNet: Attention-based deep learning framework for improved gastrointestinal disease classification," Diagnostics, vol. 13, no. 10, pp. 1753, 2023.

[4] M. Nadeem et al., "Localization and classification of gastrointestinal tract disorders using explainable AI from endoscopic images," Appl. Sci., vol. 13, no. 16, pp. 9433, 2023.

[5] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," Advances in Neural Information Processing Systems, vol. 33, pp. 6840-6851, 2020.

[6] P. Dhariwal and A. Nichol, "Diffusion models beat GANs on image synthesis," Advances in Neural Information Processing Systems, vol. 34, pp. 8780-8794, 2021.

[7] K. Pogorelov et al., "KVASIR: A multi-class image dataset for computer aided gastrointestinal disease detection," in Proc. 8th ACM Multimedia Systems Conf., Taipei, Taiwan, pp. 164-169, 2017.

[8] H. Borgli et al., "HyperKvasir: A comprehensive multi-class image and video dataset for gastrointestinal endoscopy," Sci. Data, vol. 7, pp. 283, 2020.

[9] M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in Proc. Int. Conf. Machine Learning, pp. 6105-6114, 2019.

[10] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L. C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," in Proc. IEEE Conf. Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, pp. 4510-4520, 2018.

[11] A. G. Howard et al., "MobileNets: Efficient convolutional neural networks for mobile vision applications," arXiv preprint arXiv:1704.04861, 2017.

[12] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in Proc. IEEE Conf. Computer Vision and Pattern Recognition, Honolulu, HI, USA, pp. 1251-1258, 2017.

[13] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proc. IEEE Conf. Computer Vision and Pattern Recognition, Las Vegas, NV, USA, pp. 770-778, 2016.

[14] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in Proc. Int. Conf. Medical Image Computing and Computer-Assisted Intervention, Munich, Germany, pp. 234-241, 2015.

[15] M. A. Khan et al., "A framework of saliency estimation and optimal deep learning features based gastrointestinal diseases detection and classification," Diagnostics, vol. 12, no. 11, pp. 2718, 2022.

[16] S. Mohapatra et al., "Gastrointestinal abnormality detection and classification using empirical wavelet transform and deep convolutional neural network from endoscopic images," Ain Shams Eng. J., vol. 14, no. 4, pp. 101942, 2023.

[17] J. Yogapriya et al., "Gastrointestinal tract disease classification from wireless endoscopy images using pretrained deep learning model," Comput. Math. Methods Med., vol. 2021, pp. 1-12, 2021.

[18] M. H. Al-Adhaileh et al., "Deep learning algorithms for detection and classification of gastrointestinal diseases," Complexity, vol. 2021, pp. 1-12, 2021.

[19] A. Majid et al., "Classification of stomach infections: A paradigm of convolutional neural network along with classical features fusion and selection," Microsc. Res. Tech., vol. 83, no. 5, pp. 562-576, 2020.

[20] C. Kumar and D. M. N. Mubarak, "Classification of early stages of esophageal cancer using transfer learning," IRBM, vol. 43, pp. 251-258, 2021.

[21] A. Ahmed, "Classification of gastrointestinal images based on transfer learning and denoising convolutional neural networks," in Proc. Int. Conf. Data Science and Applications, Springer, Singapore, pp. 631-639, 2022.

[22] J. Escobar, K. Sanchez, C. Hinojosa, H. Arguello, and S. Castillo, "Accurate deep learning-based gastrointestinal disease classification via transfer learning strategy," in Proc. XXIII Symp. Image, Signal Processing and Artificial Vision, Popayán, Colombia, pp. 1-5, 2021.

[23] T. Aoki et al., "Automatic detection of erosions and ulcerations in wireless capsule endoscopy images based on a deep convolutional neural network," Gastrointest. Endosc., vol. 89, pp. 357-363, 2019.

[24] S. Jain, A. Seal, A. Ojha, A. Yazidi, J. Bures, I. Tacheci, and O. Krejcar, "A deep CNN model for anomaly detection and localization in wireless capsule endoscopy images," Comput. Biol. Med., vol. 137, pp. 104789, 2021.

[25] Y. Yuan, J. Wang, B. Li, and M. Q.-H. Meng, "Saliency based ulcer detection for wireless capsule endoscopy diagnosis," IEEE Trans. Med. Imaging, vol. 34, pp. 2046-2057, 2015.

[26] J. H. Lee et al., "Spotting malignancies from gastric endoscopic images using deep learning," Surg. Endosc., vol. 33, pp. 3790-3797, 2019.

[27] D. Jha et al., "Kvasir-SEG: A segmented polyp dataset," in Proc. MultiMedia Modeling: 26th Int. Conf., Daejeon, South Korea, pp. 451-462, 2020.

[28] D. Jha et al., "ResUNet++: An advanced architecture for medical image segmentation," in Proc. IEEE Int. Symp. Multimedia, San Diego, CA, USA, pp. 225-2255, 2019.

[29] D. Jha et al., "Real-time polyp detection, localization and segmentation in colonoscopy using deep learning," IEEE Access, vol. 9, pp. 40496-40510, 2021.

[30] D. P. Fan et al., "PRANet: Parallel reverse attention network for polyp segmentation," in Proc. Int. Conf. Medical Image Computing and Computer-Assisted Intervention, Lima, Peru, pp. 263-273, 2020.

[31] B. H. M. van der Velden, H. J. Kuijf, K. G. A. Gilhuijs, and M. A. Viergever, "Explainable artificial intelligence (XAI) in deep learning-based medical image analysis," Med. Image Anal., vol. 79, pp. 102470, 2022.

[32] W. Kusakunniran et al., "COVID-19 detection and heatmap generation in chest X-ray images," J. Med. Imaging, vol. 8, no. Suppl 1, pp. 014001, 2021.

[33] K. Bae, H. Ryu, and H. Shin, "Does Adam optimizer keep close to the optimal point?" arXiv preprint arXiv:1911.00289, 2019.

[34] Y. Ho and S. Wookey, "The real-world-weight cross-entropy loss function: Modeling the costs of mislabeling," IEEE Access, vol. 8, pp. 4806-4813, 2020.

[35] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," J. Big Data, vol. 6, pp. 60, 2019.

[36] K. Weiss, T. M. Khoshgoftaar, and D. Wang, "A survey of transfer learning," J. Big Data, vol. 3, pp. 1-40, 2016.

[37] A. A. Taha and A. Hanbury, "Metrics for evaluating 3D medical image segmentation: Analysis, selection, and tool," BMC Med. Imaging, vol. 15, pp. 29, 2015.

[38] S. Seo et al., "Predicting successes and failures of clinical trials with outer product-based convolutional neural network," Front. Pharmacol., vol. 12, pp. 670670, 2021.

[39] M. N. Noor, T. A. Khan, F. Haneef, and M. I. Ramay, "Machine learning model to predict automated testing adoption," Int. J. Softw. Innov., vol. 10, pp. 1-15, 2022.

[40] M. N. Noor, M. Nazir, S. Rehman, and J. Tariq, "Sketch-recognition using pre-trained model," in Proc. National Conf. Engineering and Computing Technology, Islamabad, Pakistan, Jan. 2021.

[41] Y. Masmoudi, M. Ramzan, S. A. Khan, and M. Habib, "Optimal feature extraction and ulcer classification from WCE image data using deep learning," Soft Comput., vol. 26, pp. 7979-7992, 2022.

[42] M. Habib, M. Ramzan, and S. A. Khan, "A deep learning and handcrafted based computationally intelligent technique for effective COVID-19 detection from X-ray/CT-scan imaging," J. Grid Comput., vol. 20, pp. 23, 2022.