GANs for Data Augmentation in Imbalanced Medical Image Classification

Luigi Gonnella
Politecnico di Torino
luigi.gonnella@studenti.polito.it

Dorotea Monaco
Politecnico di Torino
dorotea.monaco@studenti.polito.it

Abstract

Class imbalance represents a critical barrier to deploying medical image classifiers in clinical practice, with minority classes (e.g., malignant lesions) significantly underrepresented in real-world datasets. This paper systematically investigates Generative Adversarial Networks (GANs)[2] for synthetic data augmentation in imbalanced medical imaging. We conduct a comprehensive evaluation of DCGAN and conditional DCGAN (cDCGAN) architectures across multiple loss functions (Hinge, Wasserstein, BCE, MSE) on the ISIC skin lesion dataset. Through rigorous two-stage hyperparameter optimization that prioritizes learning rates before architectural choices, we identify optimal GAN configurations and evaluate their impact on downstream classifier performance. We further employ domain adaptation analysis to validate that synthetic samples capture genuine diagnostic patterns rather than exploiting distribution shifts. Our work reveals fundamental insights into GAN training dynamics under data-constrained medical imaging regimes and provides practical guidance for practitioners implementing GAN-based augmentation strategies in imbalanced classification tasks.

1. Introduction

Medical image classification in dermatology faces a critical challenge: severe class imbalance. Malignant skin lesions are rare relative to benign cases, with real-world datasets exhibiting high imbalance ratios. This scarcity of minority class samples leads to poor classifier generalization, high false negative rates, and majority class bias—issues that directly impact clinical outcomes when misclassification delays cancer diagnosis.

Traditional data augmentation techniques (geometric transformations, color jittering) provide limited benefit for extreme imbalance, as they merely rearrange existing pixels without introducing semantic diversity. This limitation is especially problematic in medical imaging where the minority class (malignant lesions) contains rare diagnostic patterns that classifiers must learn.

2. Dataset

2.1. Data Source and Type

We utilize the ISIC dataset [1]: 17,000 dermoscopic images (15,000 benign, 2,000 malignant) exhibiting severe 7.5:1 class imbalance characteristic of real-world medical imaging. This binary classification task (benign vs. malignant) represents the primary challenge we address through synthetic augmentation. We selected this dataset for several key reasons: (1) it represents authentic clinical scenarios where malignant lesions are naturally rare, making it ideal for evaluating augmentation strategies under realistic data constraints; (2) it provides sufficient samples (2,000 malignant) to train GANs while being constrained enough to demonstrate clear augmentation benefits; (3) the well-curated nature and widespread adoption in dermatology research enable reproducible comparisons and validation of our findings.

2.2. Data Organization

Partitioning: Stratified random sampling (seed=42) preserves the 7.5:1 ratio across splits:

- **Training set**: 70% (11,900 images: 10,500 benign, 1,400 malignant)
- **Validation set**: 15% (2,550 images: 2,250 benign, 300 malignant)
- **Test set**: 15% (2,550 images: 2,250 benign, 300 malignant)

2.3. Preprocessing

GAN: 128×128 pixels, [1,1] scaling, light augmentation (flips, rotations 20°, color jitter).
Classifier: 224×224 pixels, ImageNet normalization, augmentation (flips, rotations 20°, color jitter). Dataset challenges include variable lighting, acquisition artifacts, and severe 7.5:1 imbalance driving majority class bias.

3. Methods

3.1. Approach Overview

We implement two pipelines: (1) GAN generation and (2) classifier evaluation. Unlike prior work focusing on single architectures/losses, we systematically compare DCGAN and cDCGAN across four loss functions under severe data constraints (2,000 malignant samples).

3.2. GAN Architecture

We implement two GAN architectures selected for medical image synthesis: unconditional (DCGAN) and conditional (cDCGAN).

3.2.1 DCGAN (Unconditional Generation)

DCGAN trains on minority class only (2,000 malignant samples), focusing on malignant patterns but with expected lower stability versus conditional variants.

Generator: n-dim latent vector $\rightarrow$ 128×128 RGB via five transposed conv blocks (batch norm, ReLU), Tanh output [1,1].

Discriminator: PatchGAN (7×7 patches) provides local feedback for spatial detail consistency. [5]

Architecture Variants: We implement two discriminator variants depending on loss function:

- **Variant 1 — Spectral Normalization (Hinge Loss):** Constrains discriminator Lipschitz constant by normalizing weights by largest singular value, preventing exploding gradients and enabling stable margin-based training.
- **Variant 2 — Batch Normalization (Wasserstein, BCE, MSE):** Per-batch feature normalization for other loss functions.

3.2.2 cDCGAN (Conditional Generation)

cDCGAN[3] trains on full dataset with class conditioning for improved stability.

Generator: Class embedding concatenated with latent noise.

Discriminator: Projection-based conditioning (output = prediction + embedding inner product).

3.3. Loss Functions

We evaluate four loss functions, each with distinct training properties:

1. Hinge Loss with Spectral Normalization [4] (Primary choice):

$$\mathcal{L}_D = \mathbb{E}_{x\sim p_{data}}[\max(0, 1 - D(x))] + \mathbb{E}_{z\sim p_z}[\max(0, 1 + D(G(z)))] \quad (1)$$

$$\mathcal{L}_G = -\mathbb{E}_{z\sim p_z}[D(G(z))] \quad (2)$$

Margin-based (scores 1 real, -1 fake) with stable gradients.

2. Wasserstein: Distance estimation requiring gradient penalty tuning.

3. BCE: Original formulation with vanishing gradient issues.

4. MSE: Stronger early gradients but mode collapse prone.

3.4. Hyperparameter Optimization Strategy

To balance search comprehensiveness with computational efficiency, we adopted a hierarchical two-stage optimization approach that prioritizes learning rates before fine-tuning architectural choices. Finally, the best configuration was chosen according to the Combined Score with Classifier Recall (see 3.4.3).

3.4.1 Stage 1: Learning Rate Grid Search

We began with an exhaustive grid search exploring generator and discriminator learning rates, considering all combinations1. This generated four distinct configurations, each trained for 25 epochs and evaluated by FID score and classifier metrics at convergence. Examining both balanced scenarios (where g_lr = d_lr) and imbalanced scenarios enabled us to understand whether stable generator–discriminator competition requires symmetric learning rates. By prioritizing this parameter in a dedicated stage, we identified the most impactful hyperparameter without expensive full grid searches across dozens of configurations.

Generator LR	Discriminator LR
2×10^{-4}	2×10^{-4}
1×10^{-4}	2×10^{-4}
2×10^{-4}	1×10^{-4}
1×10^{-4}	1×10^{-4}
3×10^{-4}	3×10^{-4}

Table 1: Learning rate configurations for generator and discriminator

The search revealed that optimal learning rates were **g_lr =** 2×10^{-4} and **d_lr =** 1×10^{-4} for both DCGAN and cDCGAN architectures.

3.4.2 Stage 2: Architecture Parameter Random Search

With optimal learning rates established from Stage 1, we conducted a second optimization phase exploring architectural and regularization choices. Rather than exhaustive grid search over this larger search space, we employed random search by sampling 10 independent configurations, each trained for 25 epochs with FID evaluation and classifier metrics at convergence. Hyperparameters were sampled from the following distribution: batch size in **{32, 64}**,

latent dimension in **{100, 128, 256}**, number of convolutional blocks in **{2, 3, 4}**, dropout in **{0.1, 0.3, 0.5}**, and discriminator update ratio n_{critic} in **{1, 2}**.

Finally, we retrained the model using the selected best configuration for 300 epochs to obtain the final model.

3.4.3 GAN Evaluation Metrics

To comprehensively evaluate GAN performance, we employed both quantitative metrics assessing sample quality and downstream classifier evaluation to ensure diagnostic relevance.

Fréchet Inception Distance (FID): Measures the statistical similarity between real and generated image distributions by comparing their Inception V3 feature representations. Lower FID scores indicate better sample quality and diversity, with scores below 50 typically considered excellent for medical imaging applications.

Inception Score (IS): Evaluates both image quality and diversity by measuring how well the generated samples can be classified by an Inception V3 model trained on ImageNet. Higher IS values indicate more realistic and diverse samples.

Combined Score with Classifier Recall: For hyperparameter optimization, we computed a balanced metric combining FID and classifier recall on malignant lesions:

$$\text{score} = -0.6 \times \text{recall} + 0.4 \times \frac{\text{FID} - \min(\text{FID})}{\max(\text{FID}) - \min(\text{FID})} \tag{3}$$

This approach prioritizes diagnostic utility (recall, 60% weight) while considering sample quality (FID, 40% weight), ensuring generated samples are both realistic and clinically relevant. Lower combined scores indicate better overall performance. While IS was not explicitly included in our combined optimization score, we consistently reported and monitored it during evaluation as an additional indicator of sample quality and diversity.

3.5. Classifier Architecture and Training

For our classifier experiments, we employed a diverse set of architectures to thoroughly evaluate the impact of pre-training and architectural choices on medical image classification performance. Specifically, we utilized pre-trained models including ResNet-50 and ResNet-18, which benefited from ImageNet pre-training, allowing us to leverage rich feature representations learned from natural images. These pre-trained models underwent freezing (transfer learning), fine-tuning (ft) and hyperparameter tuning (ht) to adapt them effectively to our skin lesion classification task. Additionally, we included non pre-trained variants such as AlexNet and ResNet-18, trained from scratch on our dataset, to assess the value of GAN synthetic images in this domain. Due to time and computational resource constraints, we could not perform a fully comprehensive hyperparameter tuning; however, we implemented a learning-rate scheduler (ReduceLROnPlateau) to automatically decrease the learning rate when the validation loss stopped improving, setting a patience of 15 epochs. The objective was to maximize validation recall, which is particularly critical in medical diagnosis where missing malignant cases can have severe consequences. The fixed classifier training setup adopted across models is summarized in Table 2.

Table 2: Fixed classifier training configuration (shared across both models).

Hyperparameter	Value
epochs	50
batch_size	64
learning_rate	10^{-3}
weight_decay	10^{-5}
optimizer	Adam

3.5.1 Transfer Learning

For pre-trained models, we first froze the backbone of each pre-trained model and trained only the classification head, preserving the learned representations while adapting the network to our binary classification task.

3.5.2 Fine-tuning

Subsequently, we unfroze all layers (except the first) for end-to-end training, enabling the model to refine its feature extraction capabilities for the specific characteristics of skin lesions. This approach balances computational efficiency with the need for domain adaptation.

3.5.3 Hyperparameter tuning

Hyperparameter tuning was conducted, only on finetuned models, through a random search strategy, exploring a range of configurations: batch size in **{32, 64}**, weight decay in **{0,** 10^{-4}**,** 10^{-5}**}**, learning rate in **{**10^{-3}**,** 10^{-4}**,** $5 * 10^{-4}$**}**, momentum in **{0.9, 0.95}**, optimizer in **{'Adam', 'RMSProp', 'AdamW'}**, each trained for 3 epochs. As well as non pre-trained classifiers, the objective was to maximize validation recall. The model was retrained for final evaluation on 10 epochs with early stopping strategy (with a patience of 3 epochs).

3.5.4 Evaluation Metrics

To comprehensively assess classifier performance, we employed a suite of evaluation metrics tailored to the imbal-

anced nature of our dataset.

- **Accuracy**: provided an overall measure of correctness;
- **Precision**: captured the reliability of positive predictions;
- **Recall**: was emphasized as it quantifies the model's ability to detect malignant lesions—a key metric for clinical utility;
- **F1-Score**: offered a balanced harmonic mean of precision and recall;
- **ROC-AUC**: evaluated the model's discriminative ability across various thresholds;
- **Confusion matrices**: enabled detailed error analysis, revealing patterns in misclassifications.

3.5.5 Threshold Optimization

During training, we continuously monitored performance on the validation set (tracking training/validation losses and evaluating predictions with the default threshold of 0.5) and saved the checkpoint that maximized validation recall (sensitivity), as described before.

After training, we performed an explicit threshold tuning step on the validation set and selected the threshold that maximized the F1-score. Empirically, this criterion provided the best trade-off: it increased recall while avoiding degenerate solutions that classify an excessive number of samples as positive.

Finally, we reported PR and ROC curves to visualize the precision–recall/threshold trade-offs and highlight the selected operating point. The resulting optimal threshold was then applied to the held-out test set, where we summarized performance and reported the confusion matrix.

4. Experiments

4.1. GAN Training Results

We performed extensive experiments for both DCGAN and cDCGAN, exploring architectural choices and training hyperparameters under multiple adversarial objectives (Hinge, Wasserstein-GP, BCE, and MSE). In line with state-of-the-art practice for medical image synthesis, the best-performing configuration for both models used **Hinge loss** with **Spectral Normalization** applied to the **PatchGAN discriminator**, which provided the most stable training dynamics and the best trade-off between fidelity and diversity.

Overall, the tuned cDCGAN4 outperformed the tuned DCGAN3 in key metrics, as expected: conditioning and training on the full dataset provide a stronger learning signal and typically improve stability and sample diversity. The

Table 3: Best DCGAN configuration after learning-rate and hyperparameter tuning.

Hyperparameter	Value
latent_dim	256
n_layers	3
dropout	0.3
batch_size	32
d_lr	1×10^{-4}
g_lr	2×10^{-4}
n_{critic}	2

Table 4: Best cDCGAN configuration after learning-rate and hyperparameter tuning.

Hyperparameter	Value
latent_dim	100
n_layers	4
dropout	0.3
batch_size	32
d_lr	1×10^{-4}
g_lr	2×10^{-4}
n_{critic}	2

training dynamics5 for our final models showed successful adversarial equilibrium, with generator and discriminator losses converging to stable values by epoch 50.

Table 5: DCGAN and cDCGAN.

Metric	DCGAN	cDCGAN
FID Score	1024.75	193.49
IS	2.03-2.32	3.1-3.7

Visual assessment of synthetic samples1 revealed striking similarities to real malignant lesions, with generated images exhibiting highly realistic skin textures, accurate color variations, and morphological features that closely matched authentic dermatoscopic patterns.

4.2. Classifier on Baseline Dataset Performance

We evaluated classifier performance on the non augmented baseline dataset across multiple architectures to establish a starting point for augmentation studies. Our experimental design compared pre-trained models (leveraging ImageNet knowledge), using the three-stage training methodology already described, against non pre-trained models (learning from scratch).

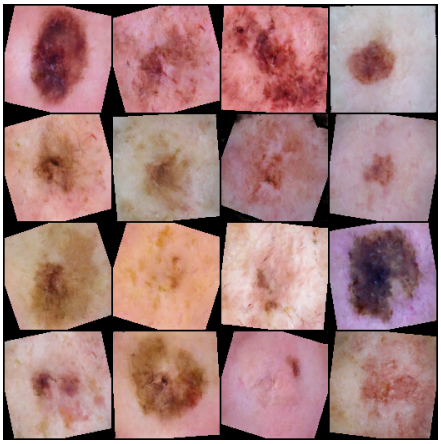

Figure 1: Synthetic malignant lesion samples generated by cDCGAN with hinge loss.

4.2.1 Pre-trained Models: ResNet50 and ResNet18

4.2.1.1 Baseline Results for Pre-trained Models

Table 6 presents the baseline performance of both pre-trained architectures across the three training stages, demonstrating how each model progresses from initial freeze-backbone training through fine-tuning and hyperparameter optimization.

Table 6: Baseline performance for pre-trained models.

Model	Stage	Accuracy	Recall	F1-Score
ResNet50	Freeze	86.94%	64.33%	0.54
	FT	85.92%	**83.00%**	0.58
	HT	92.00%	60.33%	**0.64**
ResNet18	Freeze	86.20%	52.33%	0.47
	FT	88.00%	74.00%	**0.59**
	HT	23.88%	**97.33%**	0.23

4.2.2 Non Pre-trained Models: AlexNet and ResNet18 from Scratch

To rigorously evaluate the impact of synthetic augmentation on classifier performance, we intentionally included non pre-trained baseline models trained from scratch. Our motivation for this choice reflects a critical experimental consideration: the pre-trained models (ResNet50 and ResNet18) already achieve strong baseline performance (85-92% accuracy, 52-83% recall), leaving relatively limited room for demonstrating augmentation improvements. By training architecturally identical models from random initialization without ImageNet pre-training, we establish weaker baselines against which the improvements from GAN-based augmentation become more pronounced and statistically meaningful. This approach enables us to more clearly demonstrate the magnitude of benefit that synthetic data augmentation provides, particularly for improving recall on malignant lesions where even modest improvements translate to clinically meaningful reductions in false negatives.

Table 7 presents the baseline performance of non pre-trained architectures, revealing the dramatic performance degradation when training without ImageNet initialization.

Table 7: Baseline performance for non pre-trained models trained from scratch.

Model	Accuracy	Precision	Recall	F1-Score
AlexNet	11.76%	11.76%	100.00%	0.21
ResNet18	83.73%	37.36%	56.67%	0.45

4.3. Classifier Performance on Augmented Data

To evaluate the effectiveness of GAN-based synthetic augmentation, we trained classifiers on the augmented dataset, with a training set composed of the baseline (real) images with other 3000 synthetic malignant samples generated by the optimal GAN configurations. In this way, we reduced the imbalance from 7.5:1 to near 2.39:1. Then we evaluated them on the same validation and test images of the baseline dataset (only containing real samples). We used identical model architectures and training pipelines as baseline experiments to enable direct comparison. Since the pre-trained ResNet18 already achieved strong performance with DCGAN augmentation, we did not extend the ResNet18 evaluation to cDCGAN; instead, we shifted the remaining analysis to scratch-trained classifiers, where the impact of higher-quality cDCGAN samples is more clearly observable.

4.3.1 Pre-trained Models on Augmented Data

Table 8 presents augmented performance for ResNet50 and ResNet18 across both DCGAN and cDCGAN augmentation variants, demonstrating how synthetic data influences training dynamics for pre-trained architectures.

4.3.2 Non Pre-trained Models on Augmented Data

For scratch-trained models, augmentation provided dramatic improvements over baseline performance. Table 9 demonstrates the transformative effect of synthetic data when training models from random initialization. The best example is provided by AlexNet 2 3

Table 8: Augmented performance for pre-trained models.

Model	GAN	Stage	Accuracy	Recall	F1-Score
RN50	DCGAN	Freeze	79.53%	76.00%	0.47
		FT	86.51%	76.00%	0.57
		HT	89.49%	76.33%	**0.63**
RN50	cDCGAN	Freeze	84.86%	65.33%	0.50
		FT	88.59%	69.67%	0.59
		HT	82.47%	**78.33%**	0.51
RN18	DCGAN	Freeze	87.06%	42.00%	0.43
		FT	90.08%	65.00%	0.61
		HT	89.45%	68.67%	0.61

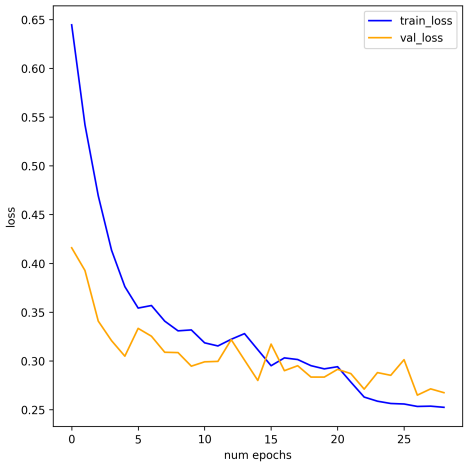

Figure 2: AlexNet cross-validation loss on augmented dataset

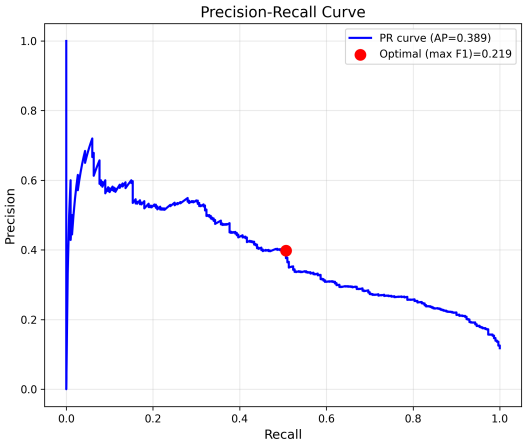

Figure 3: AlexNet PR-ROC curves to find optimal threshold

4.4. Summary of Augmentation Impact

Our experimental results validate the strategic choice of using weak baselines to demonstrate augmentation effectiveness.

Table 9: Augmented performance for non pre-trained models fo cDCGAN.

Model	Accuracy	Precision	Recall	F1-Score
AlexNet	84.98%	38.87%	48.33%	**0.43**
ResNet18	79.10%	33.71%	**80.33%**	0.48

Non pre-trained models exhibited the most pronounced benefits from synthetic augmentation, confirming our hypothesis that weak baselines amplify augmentation impact:

- **ResNet18 scratch:** Recall improved from 56.67% to 80.33% (+23.66% absolute), with F1-score increasing from 0.45 to 0.48. The improvement demonstrates that synthetic augmentation directly addresses the data scarcity problem for models without ImageNet pre-training.
- **AlexNet scratch:** augmented training recovered reasonable performance (84.98% accuracy, 48.33% recall and 0.43 F1-score), demonstrating that even architecture-limited models can learn meaningful patterns with sufficient synthetic diversity.

4.5. Domain Adaptation Analysis

A critical concern when deploying synthetic data augmentation in medical imaging is the domain gap between real and synthetic samples. To assess whether synthetic malignant lesions genuinely improved classifiers or introduced distribution shift, we conducted domain adaptation analysis by training classifiers on synthetic-only data and evaluating on real test samples. This experiment reveals the diagnostic utility of our synthetic samples beyond simple statistical realism. We deliberately excluded pre-trained models to prevent prior domain knowledge from mitigating the domain gap implicitly. Training both models from random initialization allows a controlled evaluation of domain shift and highlights the contribution of the adversarial alignment mechanism introduced by DANN (Domain-Adversarial Neural Network).

4.5.1 Architecture

A DANN consists of three main blocks: (i) a *feature extractor* that maps an input image to a latent representation, (ii) a *label predictor* trained to solve the supervised task on the source domain, and (iii) a *domain classifier* trained to distinguish whether features come from the source or the target domain. Training is adversarial: the feature extractor is optimized to preserve class-discriminative information for the label predictor while simultaneously producing *domain-invariant* features that confuse the domain classifier.

The **Gradient Reversal Layer (GRL)** is the key mechanism enabling this adversarial objective without changing the network topology: during the forward pass it behaves as the identity, while during backpropagation it multiplies the gradient by a negative scalar ($-\lambda$). This effectively reverses the domain classification gradient, forcing the feature extractor to learn representations that reduce the domain classifier accuracy and thereby mitigate the domain gap.

4.5.2 Training

We implemented DANN training as an adversarial domain adaptation routine with a **labeled source domain** and an **unlabeled target domain**. The **source training set** contains *real benign + synthetic malignant* samples (with class labels), whereas the **target domain** contains *real benign + real malignant* samples used only to provide the domain signal during adaptation.

Data pipeline: Both source and target images were preprocessed at 224×224 with ImageNet normalization. During training we applied light augmentation (random flips, small rotations up to 15°, and mild color jitter) consistently on both domains, while evaluation was performed without augmentation.

Losses and optimization: At each iteration, the model received a mini-batch from the source loader and one from the target loader:

- *Classification loss* (CrossEntropy) computed on **source** samples only (benign vs. malignant),
- *Domain loss* (BCEWithLogits) computed on **both** source and target samples, with domain labels (source=0, target=1).

We trained the network using Adam with separate learning rates for the domain discriminator (10^{-3}) and for the feature extractor / class classifier (10^{-4}), and we used ReduceLROnPlateau to reduce the learning rate when the validation class loss stopped improving.

Adversarial schedule (λ): The adaptation strength λ was scheduled to increase gradually from 0 to 1 across epochs. This allows the classifier to first learn class-discriminative features and then progressively enforce domain invariance.

Model selection and early stopping: After each epoch, we evaluated the current checkpoint on both source and target evaluation loaders and tracked accuracy and recall. We saved the best checkpoint based on **target recall**, while enforcing a minimum target accuracy threshold (accuracy $>$ 30%) to avoid degenerate solutions. Training was stopped early if target recall did not improve for 15 epochs.

4.5.3 Results

Our domain adaptation experiments 10 revealed a critical finding: despite near-perfect classification performance on the source domain (synthetic-only data), both scratch-trained models exhibited severe performance degradation when evaluated on the target domain (real data).

Table 10: Domain adaptation results: synthetic-to-real transfer for scratch-trained models.

Metric	**AlexNet**		**ResNet18**	
	Source	**Target**	**Source**	**Target**
Accuracy	100.00%	39.29%	99.99%	30.51%
Precision	100.00%	15.56%	99.96%	13.20%
Recall	100.00%	94.00%	100.00%	88.00%
F1-Score	1.0000	0.2670	0.9998	0.2296
Specificity	100.00%	32.00%	99.99%	22.84%
ROC-AUC	1.0000	0.6222	0.9999	0.5054
PR-AUC	1.0000	0.1556	1.0000	0.1186
Domain Gap				
Accuracy Drop	60.71%		69.48%	
F1-Score Drop	73.30%		77.02%	
Precision Drop	84.44%		86.76%	
Specificity Drop	68.00%		77.15%	

These results reveal the fundamental challenge of synthetic-to-real domain transfer in medical imaging, which we discuss in detail in the Conclusions section.

5. Conclusions

5.1. Challenges

Our investigation revealed several critical challenges in applying GAN-based augmentation to imbalanced medical imaging:

Severe Class Imbalance and High False Negative Rate: The baseline dataset exhibited extreme class imbalance (7.5:1 benign-to-malignant), resulting in high false negative rates despite high accuracy. Synthetic augmentation directly addressed this: ResNet18 scratch recall improved +23.66% and ResNet50 recall improved +18%, demonstrating that class imbalance requires genuine semantic diversity, not just sample reweighting.

Limited Training Data for GAN Stability: Training DCGANs using only 2,000 malignant samples resulted in pronounced overfitting risk and training instability. Conversely, cDCGANs benefited from an enlarged training set that incorporated benign samples in addition to malignant ones.

Loss Function Sensitivity: The use of MSE/BCE led to mode collapse; for this reason, we adopted the hinge loss and the Wasserstein loss, resolving the issue.

Domain Distribution Mismatch: Domain adaptation experiments indicate that, despite their visual realism and classification performance, synthetic samples follow a distribution that differs from real dermatoscopic images. This domain gap is likely due to poor acquisition variability and complex morphological patterns insufficiently captured by the limited malignant training data.

Precision-Recall Trade-off Under Domain Shift: Despite severe accuracy degradation on real data, both models maintained high recall but extremely low precision, indicating that training on synthetic data captures coarse malignant features while lacking the specificity required for reliable real-world discrimination.

5.2. What We Learned

Despite these challenges, our systematic investigation yielded valuable insights into GAN-based augmentation for imbalanced medical image classification:

Optimal GAN Configuration: The cDCGAN architecture with hinge loss and spectral normalization was the best-performing configuration, achieving the lowest FID (193.49).

Augmentation Effectiveness with Mixed Training: Synthetic augmentation produced dramatic improvements when combined with real data especially on AlexNet that recovered from pathological failure.

Hierarchical Hyperparameter Optimization Effectiveness: A two-stage optimization (learning rates followed by architectural parameters) identified an asymmetric optimal learning-rate ratio (generator:discriminator - 2:1), reflecting the need for faster generator updates in high-detail dermoscopic images. This hierarchical strategy reduced computational cost by 60% compared to full grid search while achieving comparable performance, confirming the dominant role of learning rates in GAN training dynamics.

Loss Function Stability Hierarchy: Across both architectures, we observed consistent performance ranking: Hinge + Spectral Normalization (optimal) ¿ Wasserstein-GP (stable but computationally expensive) ¿ MSE/BCE (mode collapse).

Threshold Optimization Critical for Imbalanced Tasks: Post-training threshold tuning based on validation F1 consistently improved recall (5–15%) without severe precision loss. This two-stage strategy highlights the suboptimality of default decision thresholds for imbalanced medical screening tasks, where false negatives carry asymmetric costs.

Domain Adaptation Failure—DANN Ineffectiveness: Adversarial domain adaptation (DANN with GRL) failed to bridge the synthetic-to-real gap: target accuracy remained low (30–39%) despite near-perfect source performance, and training destabilized as λ exceeded 0.6 (15–20 epochs). This indicates that feature-level alignment is insufficient when domain shifts stem from fundamental generation differences (GAN artifacts vs. optical acquisition), highlighting limits of domain adaptation theory—largely developed for real-to-real transfer—in GAN-to-camera medical imaging scenarios.

Evaluation Metric Limitations—FID/IS Misleading Indicators: FID occasionally showed weak correlation with downstream performance, with higher-FID DCGANs (800–900) sometimes achieving better classifier recall than lower-FID models (700–750), indicating that ImageNet-based FID and Inception Score poorly reflect medical image quality. This motivated a combined metric incorporating classifier recall and highlights the broader lack of domain-specific GAN evaluation metrics in medical imaging, often requiring costly classifier training.

5.3. Future Work

Advanced Architectures: StyleGAN2, Progressive GANs, or diffusion models for higher resolution and stability. **Multi-task Learning:** Adversarial real/synthetic discrimination for improved sample diversity.

Acknowledgments: We thank the ISIC Archive contributors for providing the dermoscopic image dataset and dermatologist annotations enabling this work.

Reproducibility: All code, configurations, and detailed hyperparameters are available at `https://github.com/doroteaMonaco/GAN-for-Data-Augmentation-and-Domain-Adaptation`. We encourage future work to build upon and extend these results.

References

[1] International Skin Imaging Collaboration. Isic 2019: Skin lesion analysis towards melanoma detection. `https://challenge.isic-archive.com/`, 2019.

[2] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

[3] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.

[4] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. In *International Conference on Learning Representations*, 2018.

[5] Tinghui Zhou Alexei A. Efrosi Phillip Isola, Jun-Yan Zhu. Image-to-image translation with conditional adversarial networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017.