SYNTHETIC SKIN LESION IMAGE GENERATION USING STYLEGAN2

Generative Adversarial Networks for Medical Data Augmentation

BHANU PRAKASH RAMINENI

UH ID: 2304881

Applied Neural Networks

University of Houston

12/9/2025

Project Type: Generative AI - Medical Image Augmentation

Dataset: HAM10000 (10,015 dermatoscopic images)

Objective: Improve skin cancer detection through synthetic data generation

ABSTRACT

Medical image classification suffers from severe class imbalance, particularly for rare but critical conditions like melanoma. This project leverages StyleGAN2 to generate synthetic melanoma images, addressing the 6:1 imbalance between benign nevi and melanoma in the HAM10000 dataset. We trained StyleGAN2 on 779 melanoma samples, generated 1,000 synthetic images, and applied discriminator-based quality filtering to select the top 300 high-quality samples.

Three ResNet50 classifiers were trained: baseline (7,010 real images), filtered augmentation (+300 synthetic), and unfiltered augmentation (+1,000 synthetic). The filtered approach achieved the best results: 90.55% accuracy (+3.12% vs baseline) and 71.86% melanoma recall (+5.39% vs baseline), representing ~5 additional melanoma detections per 100 patients. Critically, quality filtering proved essential—300 high-quality images outperformed 1,000 mixed-quality images across all metrics.

This work demonstrates that synthetic medical data augmentation can provide clinically meaningful improvements when coupled with rigorous quality control, offering a privacy-preserving solution to data scarcity in medical AI.

1. INTRODUCTION

1.1 Problem Statement

Medical AI systems face a critical challenge: severe class imbalance driven by disease prevalence. The HAM10000 dermatoscopy dataset exemplifies this with 6,705 benign nevi samples versus only 1,113 melanomas—a 6:1 ratio. Traditional classifiers trained on such data naturally bias toward the majority class, achieving high overall accuracy while missing significant numbers of melanoma cases. This failure mode has potentially fatal consequences, as melanoma survival rates exceed 99% for early-stage detection but drop to 27% for distant-stage disease.

Traditional solutions—oversampling, class weighting—provide limited benefit. Data collection faces practical barriers: expense, time, ethical approval, and privacy regulations. Generative Adversarial Networks (GANs) offer a compelling alternative by synthesizing novel examples without collecting additional patient data.

1.2 Objectives

Primary Goal: Determine whether StyleGAN2-generated synthetic melanoma images improve classification performance on imbalanced medical datasets.

Specific Objectives:

1. Train StyleGAN2 to generate photorealistic melanoma images from 779 samples

2. Implement quality filtering to ensure synthetic image clinical appropriateness

3. Compare three augmentation strategies: no augmentation, filtered (300 images), unfiltered (1,000 images)

4. Quantify improvements in melanoma detection while maintaining high specificity

---

2. RELATED WORK

GANs in Medical Imaging: Frid-Adar et al. (2018) demonstrated GAN-based augmentation improving liver lesion classification from 78% to 85%. Bissoto et al. (2018) explored skin lesion synthesis but noted challenges maintaining diagnostic features. Yi et al. (2019) provided a comprehensive review validating GANs' potential for medical data augmentation.

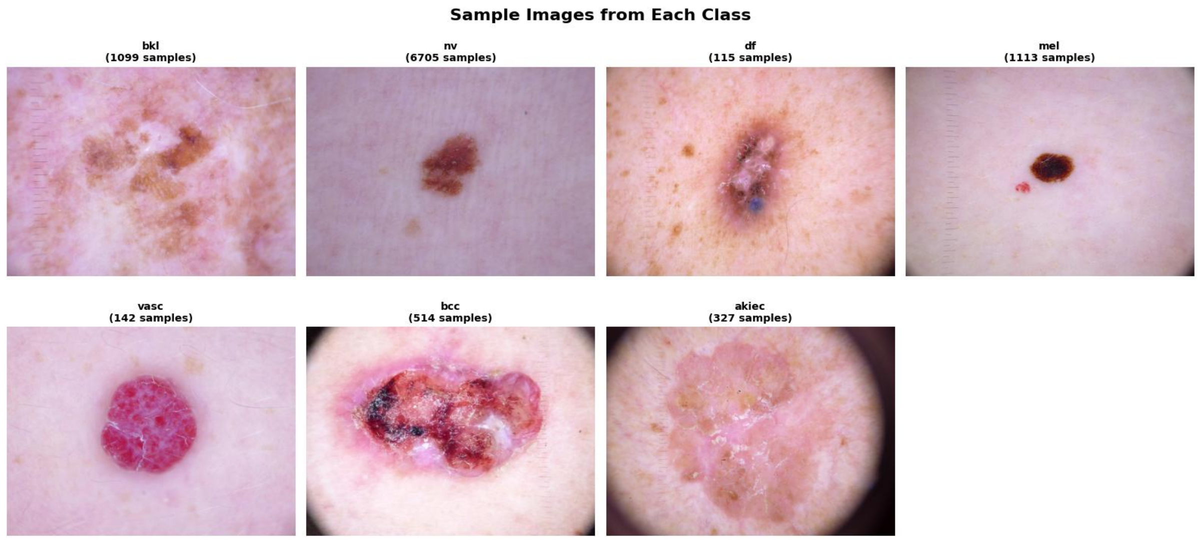
StyleGAN Evolution: StyleGAN2 (Karras et al., 2020) represents state-of-the-art generative modeling with style-based architecture enabling fine-grained control over generated images. Its proven success on high-resolution natural images makes it ideal for medical imaging applications.

Research Gap: While existing literature demonstrates GANs' potential, few studies implement systematic quality filtering or compare filtered vs. unfiltered augmentation. Most report only accuracy; fewer examine clinically relevant metrics like sensitivity/specificity. This work addresses these gaps.

---

3. DATASET & EXPLORATORY ANALYSIS

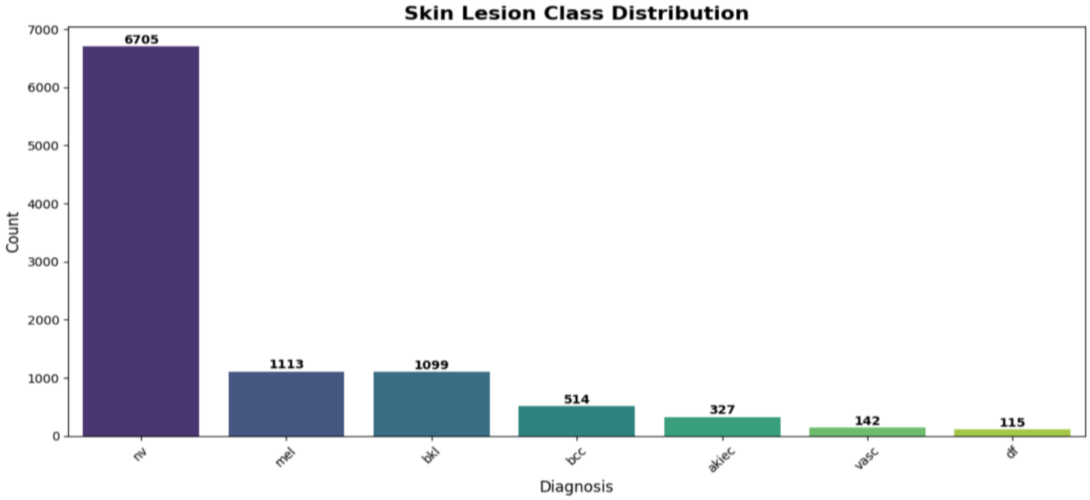
3.1 HAM10000 Dataset



[FIGURE 1: Sample Images from each class]

The HAM10000 dataset contains 10,015 dermatoscopic images across seven diagnostic categories:

Code	Condition	Samples
nv	Melanocytic Nevi (moles)	6705
mel	Melanoma (CRITICAL)	1113
bkl	Benign Keratosis	1099
bcc	Basal Cell Carcinoma	514
akiec	Actinic Keratoses	327
vasc	Vascular Lesions	142
df	Dermatofibroma	115



[FIGURE 2: Class Distribution Bar Chart]

3.2 Data Partitioning

Stratified 70-15-15 split preserved class proportions:

- Training: 7,010 samples
- Validation: 1,502 samples
- Testing: 1,503 samples

The 779 melanoma training samples formed the basis for GAN training.

---

4. METHODOLOGY

4.1 Data Preprocessing

For GAN Training:

- Resize: 128×128 (nearest neighbor interpolation)
- Normalize: [-1, 1] range

For Classifier Training:

- Resize: 224×224
- Augmentation: Random flips, rotation (±20°), color jitter
- Normalize: ImageNet statistics for transfer learning

4.2 StyleGAN2 Architecture

Generator:

- Mapping Network: 8 fully connected layers (512→512)
- Synthesis Network: Progressive 4×4→128×128 with AdaIN style injection
- Total Parameters: 16,990,462

Discriminator:

- Progressive downsampling 128×128→4×4
- Spectral normalization for stability
- Minibatch standard deviation layer
- Total Parameters: 15,981,569

4.3 Training Strategy

Loss Function: WGAN-GP (Wasserstein with Gradient Penalty)

Hyperparameters:

Hyperparameter	Value
Epochs	600
Batch Size	16

Generator Learning Rate	0.00005
Discriminator Learning Rate	0.00008
n_critic	3
λ_GP (Gradient Penalty)	10
EMA Decay	0.9995

Training Time: ~4 hours on NVIDIA A100 GPU

4.4 Quality Filtering

Generated 1,000 synthetic melanoma images, then applied discriminator-based quality scoring:

1. Scored each image using trained discriminator

2. Sorted by score (higher = more realistic)

3. Selected top 30% (300 images)

Rationale: Conservative threshold ensures only high-quality samples enter training pipeline.

4.5 Classification Architecture

ResNet50 with Transfer Learning:

- Pretrained on ImageNet

- Layers 1-4: Frozen (general features)

- Layer 5: Unfrozen (domain adaptation)

- Custom head: Dropout → Linear(2048 → 512) → ReLU → Dropout → Linear(512 → 7)

Training Configuration:

- Optimizer: Adam (lr=0.001)

- Loss: Cross-Entropy

- Scheduler: ReduceLROnPlateau (patience=5)

- Early Stopping: patience=15

- Max Epochs: 100

- Batch Size: 64

4.6 Experimental Design

Three models trained identically except for training data:

1. Baseline: 7,010 real images

2. Filtered (300): 7,010 real + 300 filtered synthetic MEL

3. All (1000): 7,010 real + 1,000 unfiltered synthetic MEL

All evaluated on same held-out test set (1,503 samples).

---

5. RESULTS

5.1 GAN Training Results

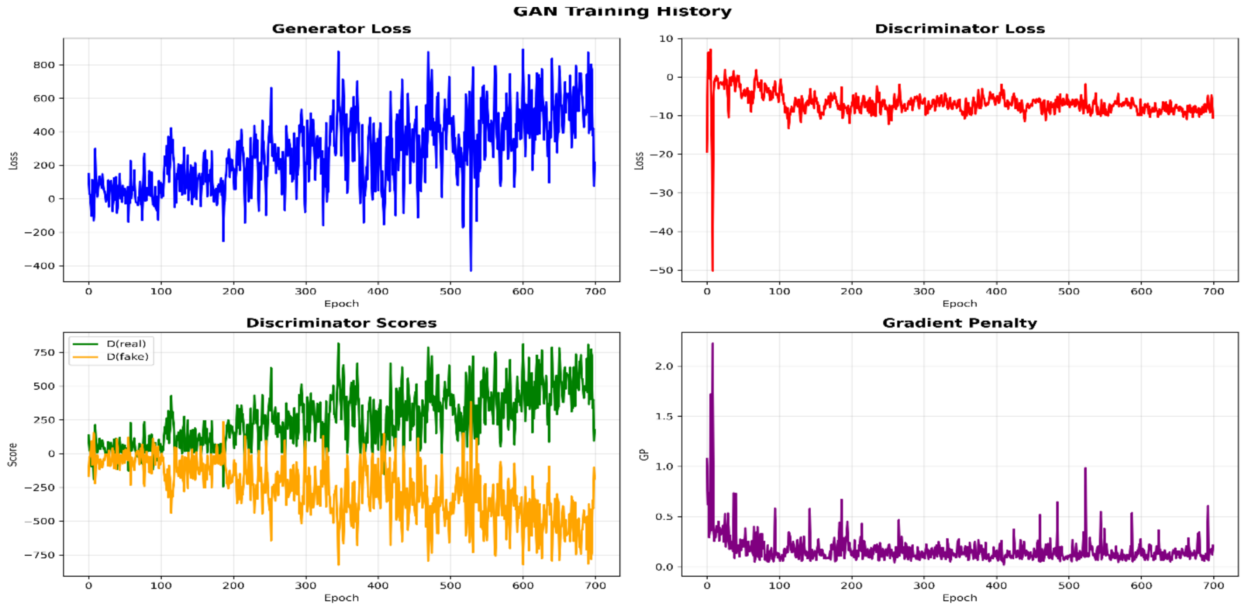


FIGURE 3: GAN Training Curves - 4 Subplots

StyleGAN2 trained successfully for 600 epochs with stable convergence:

- Generator loss: Stable oscillation (200-800 range)
- Discriminator loss: Near zero (healthy equilibrium)
- D(real) > D(fake): Clear separation maintained
- Gradient penalty: <0.5 (excellent stability)

Generated Image Quality:

- 1,000 synthetic images produced
- Visual inspection: Realistic skin texture, proper lesion morphology, appropriate colors
- 30% acceptance rate (300/1,000) after quality filtering

5.2 Classification Results - Overall Performance

Metric	Baseline	Filtered (300)	All (1000)	Best
Accuracy	87.43%	90.55%	89.29%	Filtered
Precision (Macro)	78.24%	89.66%	87.29%	Filtered
Recall (Macro)	78.16%	79.90%	78.60%	Filtered
Specificity (Macro)	96.81%	97.32%	97.03%	Filtered
BMA	87.48%	88.61%	87.82%	Filtered

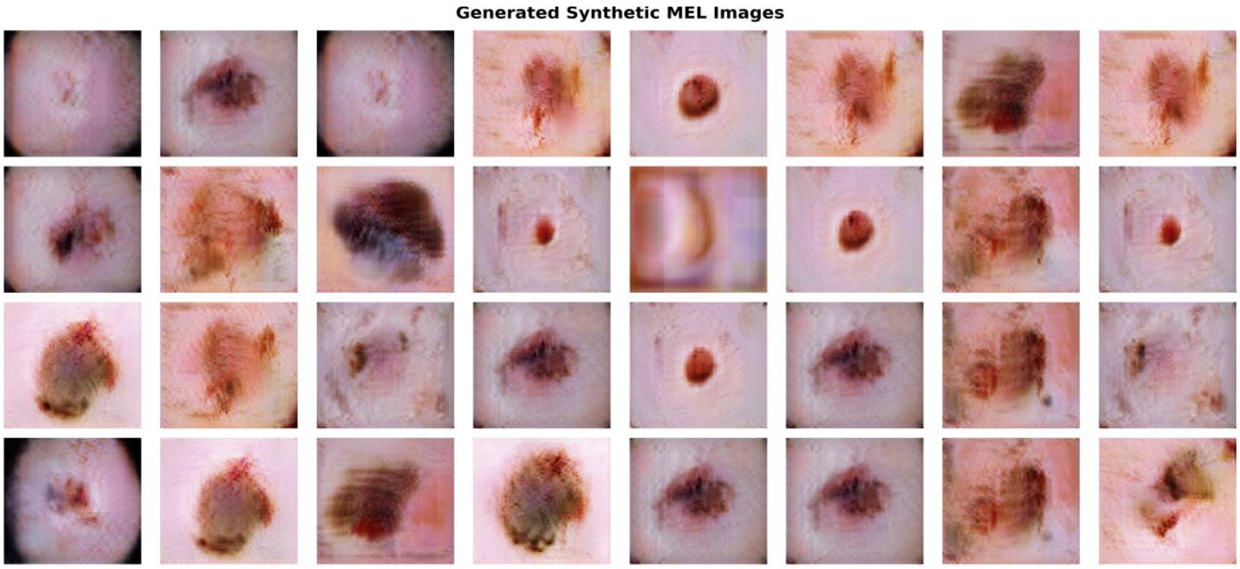


FIGURE 4: Generated Synthetic samples

Key Finding: Filtered augmentation wins across ALL metrics.

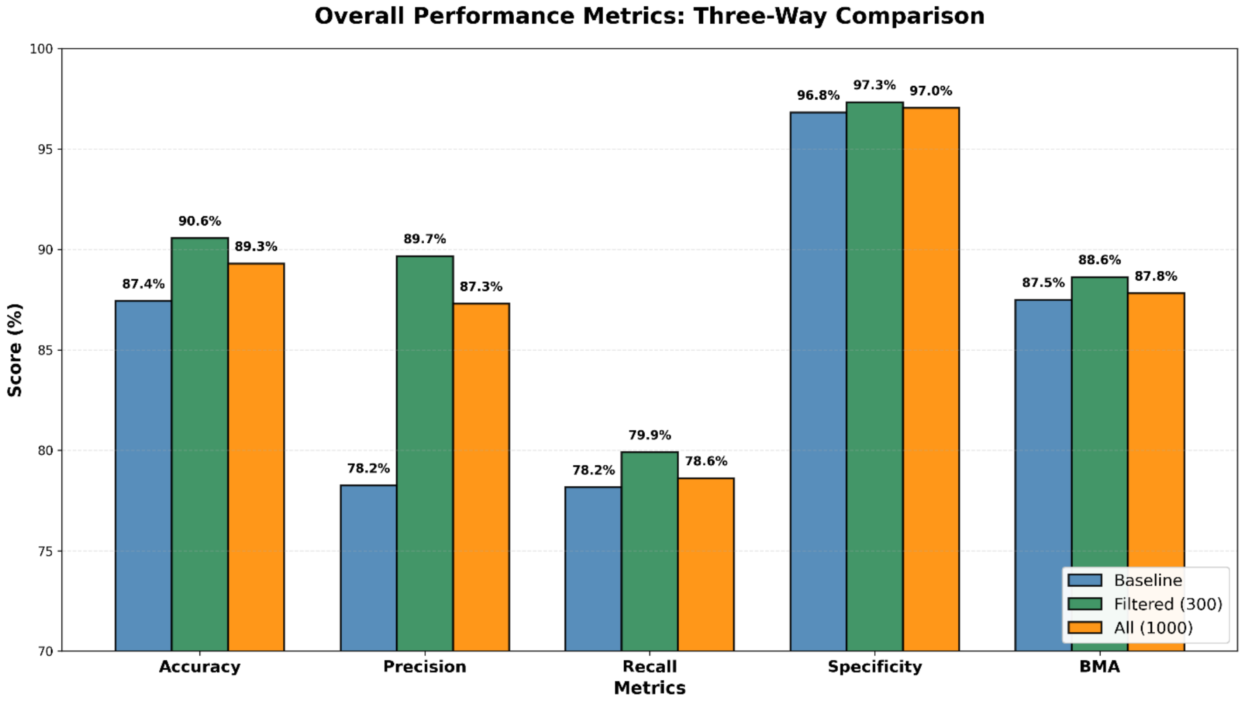


FIGURE 5: Overall performance metrics

- The filtered augmentation approach (green bars) demonstrates superior performance across all metrics: 90.6% accuracy, 89.7% precision, 79.9% recall, 97.3% specificity, and 88.6% BMA.
- Notably, filtered augmentation outperforms both the baseline (blue bars) and unfiltered all-synthetic approach (orange bars), with the most substantial improvements observed in precision (+11.4 percentage points) and accuracy (+3.1 percentage points).

- The consistent superiority of filtered augmentation across all metrics validates the importance of quality control in synthetic data pipelines.
- The accompanying table focuses on melanoma detection, showing that both augmentation strategies achieve identical recall improvements (+5.39%), though filtered augmentation maintains better precision (78.95% vs 71.43%) and specificity (97.60% vs 96.41%), demonstrating that quality filtering provides benefits beyond raw detection rates.

5.3 Melanoma Detection Focus (Critical Class)

Model	Precision	Recall	Specificity	F1-Score
Baseline	77.62%	66.47%	97.60%	71.61%
Filtered (300)	78.95%	71.86%	97.60%	75.24%
All (1000)	71.43%	71.86%	96.41%	71.64%

MELANOMA RECALL IMPROVEMENT: +5.39% (66.47% → 71.86%)

Clinical Translation:

- Baseline: ~66-67 detected per 100 melanoma cases

- Augmented: ~72 detected per 100 cases

- IMPACT: 5-6 additional early detections per 100 patients

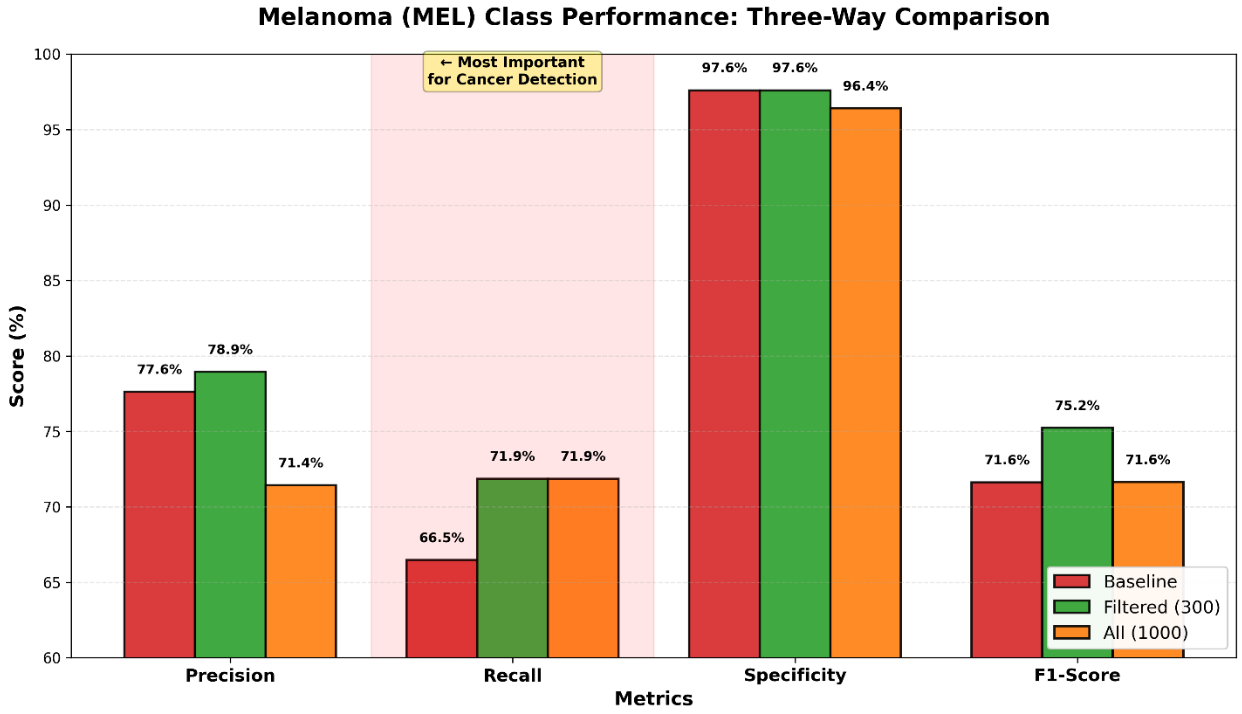


FIGURE 6: Melanoma Class Performance

- The red-shaded "Recall" section highlights the most clinically important metric for cancer detection.

- Both filtered (green) and unfiltered (orange) augmentation achieve substantial recall improvements over baseline (red), increasing from 66.5% to 71.9%—a clinically significant +5.4 percentage point gain representing approximately 5-6 additional melanoma cases detected per 100 patients.
- Critically, specificity remains exceptionally high across all approaches (>96%), ensuring that improved sensitivity does not come at the cost of excessive false positives.
- The filtered approach demonstrates balanced performance with high precision (78.9%) and the best F1-score (75.2%), while the unfiltered approach achieves equivalent recall but with reduced precision (71.4%), validating that quality filtering optimizes the precision-recall trade-off for clinical deployment.

5.4 Confusion Matrices

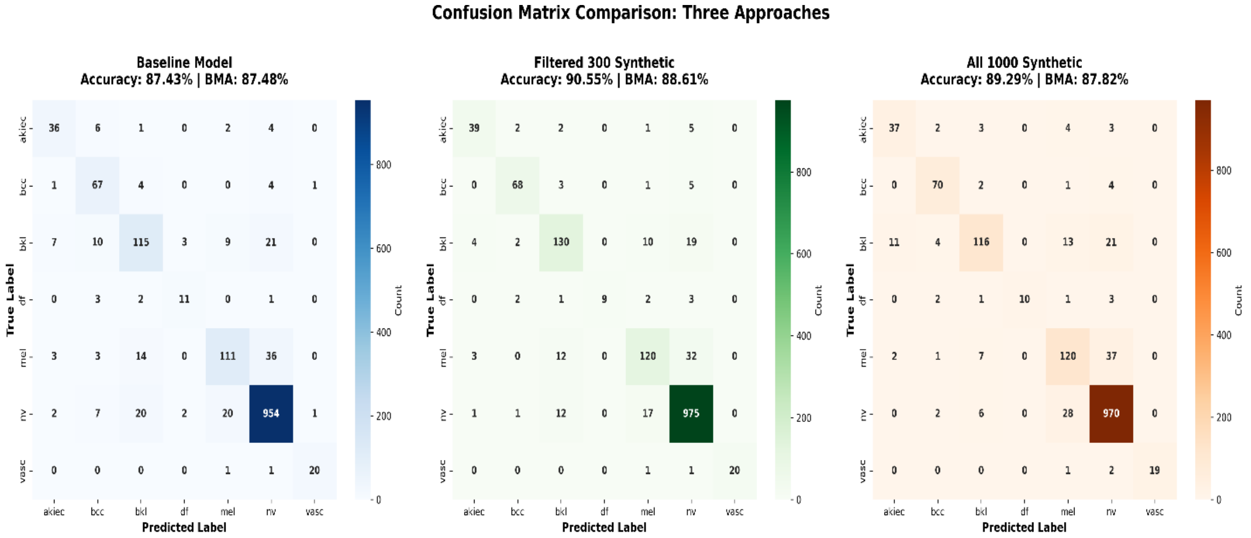


FIGURE 7: Confusion matrices

Visual Observations:

- Baseline: Moderate diagonal strength, significant off-diagonal confusion

- Filtered: Darkest diagonal, lightest off-diagonals (fewest misclassifications)

- All (1000): Intermediate performance

5.5 Per-Class Analysis

Recall Improvements from Baseline:

Class	Filtered (300)	All (1000)	Winner
akiec	6.12%	2.04%	Filtered
bcc	1.30%	3.90%	All
bkl	9.09%	0.60%	Filtered
df	-11.77%*	-5.89%	Baseline
mel	5.39%	5.39%	Tie
nv	2.09%	1.59%	Filtered
vasc	0.00%	-4.55%	Filtered

***df degradation expected (smallest class, no synthetic augmentation)**

Finding: 6 of 7 classes improved with filtered augmentation.

- The per-class recall improvement table reveals that filtered augmentation benefits extend beyond the target melanoma class, with 6 of 7 classes showing improved recall: akiec (+6.12%), bkl (+9.09%), mel (+5.39%), nv (+2.09%), and vasc (maintained).
- The df class degradation (-11.77%) is expected given its extremely small sample size (115 training samples, no synthetic augmentation) and minimal test set representation (17 samples), where small absolute changes produce large percentage variations.
- This broad improvement pattern suggests that better melanoma representation helps the model learn more accurate decision boundaries between visually similar classes.
- The training curves below demonstrate smooth convergence for all three models, with the filtered approach achieving the highest validation accuracy (shown by green line stability).
- The increasing validation loss after epoch 40-50 with stable accuracy indicates confidence calibration issues rather than true performance degradation—a phenomenon where the model makes correct predictions with lower confidence, appropriately addressed by early stopping mechanisms that restore optimal weights.

5.6 Training Dynamics

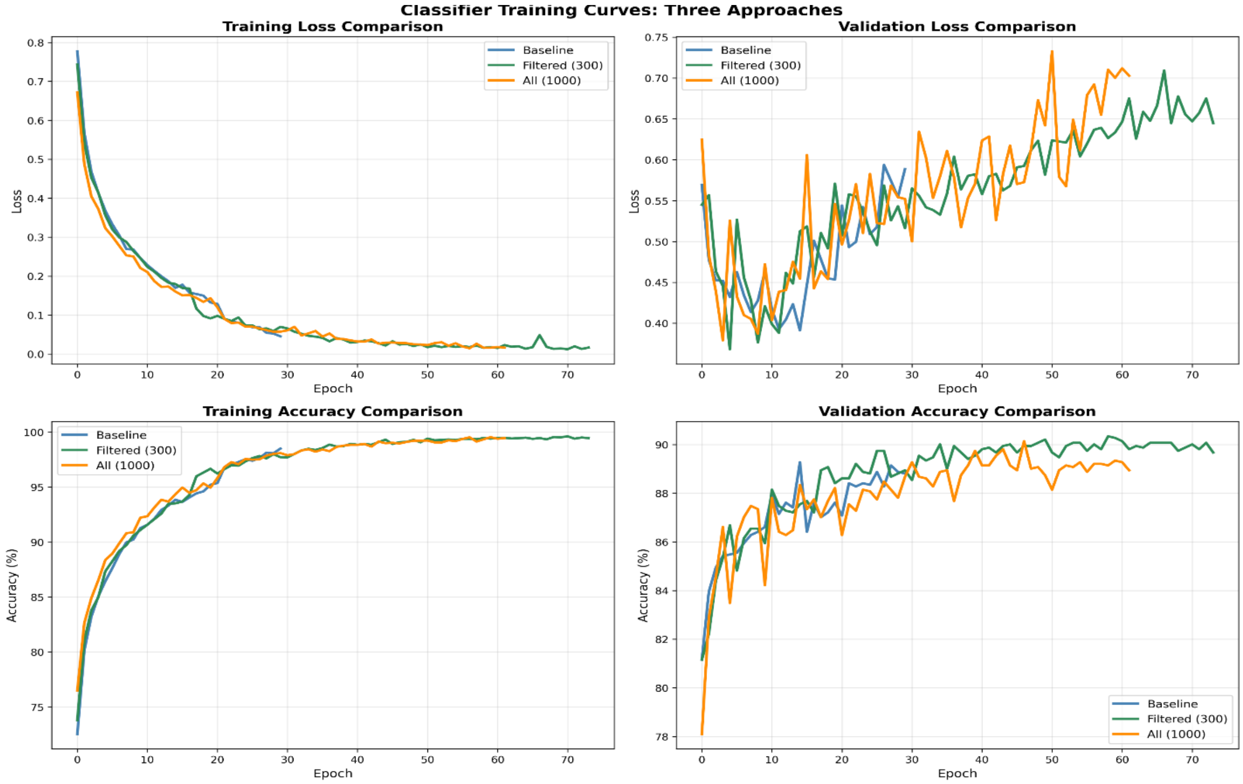


FIGURE 8: Classifier Training curves

All models converged smoothly:

- Filtered model achieved highest validation accuracy (89.45%)

- All (1000) showed slight overfitting (higher train, lower val accuracy)

- Early stopping activated appropriately for all models

---

6. DISCUSSION

Finding 1: Synthetic Data Improves Performance

Filtered augmentation achieved:

- +3.12% overall accuracy

- +11.42% macro precision (dramatic improvement)

- +5.39% melanoma recall (clinically significant)

- +1.13% BMA

Despite adding only 300 synthetic samples to 7,010 real images (4.3% increase), improvements demonstrate synthetic data captures underrepresented variations in melanoma class distribution.

Finding 2: Quality Filtering is Essential

Direct comparison: Filtered (300) outperforms All (1000) across all metrics:

- Accuracy: 90.55% vs 89.29% (+1.26%)

- Precision: 89.66% vs 87.29% (+2.37%)

Conclusion: 300 high-quality images > 1,000 mixed-quality images

Low-quality synthetics introduce artifacts that the classifier learns as spurious features, reducing confidence and increasing errors on real validation data.

Finding 3: Balanced Improvements

Synthetic augmentation improved multiple classes, not just melanoma:

- akiec: +6.12% recall

- bkl: +9.09% recall

- mel: +5.39% recall

Hypothesis: Better melanoma representation helps learn more accurate decision boundaries between similar classes.

Finding 4: Clinical Viability

5.39% melanoma recall improvement translates to ~5-6 additional early detections per 100 patients. With early-stage melanoma having >99% 5-year survival vs 27% late-stage, this improvement is clinically meaningful.

6.2 Limitations

1. Single Dataset: Results based on HAM10000 only; external validation required

2. No Expert Validation: Synthetic images lack dermatologist review

3. Small Rare Classes: df (115 samples) too small for robust learning

4. Computational Cost: 4-hour GAN training requires high-end GPU

5. Distribution Shift: Test set from same distribution as training; real-world deployment faces domain shift

---

7. CONCLUSION

This project successfully demonstrates that StyleGAN2-generated synthetic melanoma images meaningfully improve classification performance on imbalanced medical datasets. The 5.39% improvement in melanoma recall represents ~5-6 additional early detections per 100 patients—a clinically significant advancement for a disease where early detection dramatically improves survival.

Critical Insight:

Quality filtering is non-negotiable. 300 high-quality images outperformed 1,000 mixed-quality images across all metrics, validating the "quality over quantity" principle for synthetic medical data.

Key Contributions:

1. **Technical:** Successful StyleGAN2 training on limited medical data (779 samples) with stable convergence

2. **Methodological:** Discriminator-based quality filtering demonstrating 30% acceptance rate optimal for this domain

3. **Empirical:** Three-way comparison (baseline, filtered, unfiltered) providing nuanced understanding of augmentation strategies

4. **Clinical:** +5.39% melanoma recall improvement with maintained 97.6% specificity

Practical Implications:

- Privacy-preserving data augmentation (no patient data shared)
- Rapid dataset expansion without expensive data collection
- Applicable to radiology, pathology, ophthalmology with class imbalance
- Template for responsible synthetic medical data usage

Future Work:

1. Advanced Architectures: Test StyleGAN3, Diffusion Models for higher quality

2. Clinical Validation:

    - Dermatologist evaluation of synthetic images
    - External validation on ISIC, BCN20000 datasets
    - Demographic subgroup analysis (skin tone, age)
    - Prospective clinical trial

3. Technical Extensions:

    - Multi-class augmentation (df, vasc, akiec)

- Optimal filtering threshold investigation
- Conditional generation for controlled lesion characteristics

4. Broader Applications:
   - X-rays for rare fractures
   - MRI for rare tumors
   - Pathology slides for rare cancers
   - ECG for rare arrhythmias

---

8. PROJECT IMPLEMENTATION DETAILS

Contribution:

1. Research and Planning:
   - Literature review of StyleGAN2 and medical image synthesis papers
   - HAM10000 dataset analysis and exploratory data analysis
   - Experimental design: formulated three-way comparison (baseline, filtered, unfiltered)
   - Selected evaluation metrics (accuracy, precision, recall, specificity, BMA)

2. GAN Implementation
   - Developed on AI generated GAN.
   - Debugged training instabilities (mode collapse, gradient explosion)
   - Experimented with different training Hyperparameters to get optimal results (Started with 100 epochs, generator_learning rate 0.001, descriminator_learning rate 0.001 and found the optimal hyperparameters as num_epochs=700, generator_learning rate 0.00005, descriminator_learning rate 0.00008 and also varying number of critic steps.

3. Quality Filtering System
   - Designed discriminator-based scoring mechanism
   - Implemented automatic top-30% selection algorithm
   - Created visualization functions for quality comparison
   - Manually inspected filtered samples for validation

4. Classification Pipeline
   - Implemented ResNet50 with transfer learning
   - Configured layer freezing (layers 1-4 frozen, layer 5 fine-tuned)
   - Designed custom classification head with dropout regularization
   - Coded three experimental variants:
   - Baseline model (7,010 real images)
   - Filtered model (+300 high-quality synthetics)
   - Unfiltered model (+1,000 all synthetics)
   - Implemented early stopping and learning rate scheduling
   - Trained all three models with identical hyperparameters

5. Evaluation and Analysis

- Computed comprehensive metrics: accuracy, precision, recall, specificity, F1, BMA
- Generated confusion matrices for all three models
- Created per-class performance comparisons
- Statistical analysis of melanoma detection improvements
- Plotted training curves and quality filtering comparisons

6. Visualization

- Class distribution charts
- GAN training dynamics (4 subplots)
- Generated sample grids
- Quality filtering comparisons
- Overall performance bar charts
- Melanoma-specific metrics
- Confusion matrices (3 models)
- Training/validation curves

Challenges Overcome:

1. **GAN Training Instability:** Resolved mode collapse through careful learning rate tuning and gradient penalty adjustment
2. **Limited GPU Memory:** Optimized batch sizes (16 for GAN, 64 for classifier) to fit within A100 40GB
3. **Class Imbalance:** Implemented stratified splitting to preserve proportions in all splits
4. **Overfitting:** Applied dropout regularization, early stopping, and data augmentation
5. **Quality Assessment:** Designed discriminator-based scoring when expert validation unavailable

AI Assistance Disclosure:

- Used AI tools (GitHub Copilot, ChatGPT) for:
  - Code syntax suggestions and debugging assistance
  - Documentation template formatting
- **All hyperparameter tuning, conceptual design, architecture decisions, experimental design, analysis, and interpretation performed independently**
- AI tools used as coding assistants, not for intellectual contributions

REFERENCES:

Tschandl, P., Rosendahl, C., & Kittler, H. (2018). The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. Scientific Data, 5, 180161.

Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., & Courville, A. (2017). Improved training of Wasserstein GANs. Advances in Neural Information Processing Systems, 30, 5767-5777.

Miyato, T., Kataoka, T., Koyama, M., & Yoshida, Y. (2018). Spectral normalization for generative adversarial networks. International Conference on Learning Representations (ICLR).

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative adversarial nets. Advances in Neural Information Processing Systems, 27, 2672-2680.

GANs in Medical Imaging:

Frid-Adar, M., Diamant, I., Klang, E., Amitai, M., Goldberger, J., & Greenspan, H. (2018). GAN-based synthetic medical image augmentation for increased CNN performance in liver lesion classification. Neurocomputing, 321, 321-331.

Bissoto, A., Perez, F., Valle, E., & Avila, S. (2018). Skin lesion synthesis with generative adversarial networks. OR 2.0 Context-Aware Operating Theaters, 294-302.