bioengineering

MDPI

Article

Synthetic Melanoma Image Generation and Evaluation Using Generative Adversarial Networks

Pei-Yu Lin [1], Yidan Shen [2], Neville Mathew [1], Renjie Hu [3], Siyu Huang [4], Courtney M. Queen [5], Cameron E. West [6], Ana Ciurea [7] and George Zouridakis [1,2,8]*

[1] Department of Engineering Technology, University of Houston, Sugar Land, TX 77479, USA
[2] Department of Electrical and Computer Engineering, University of Houston, Houston, TX 77204, USA
[3] Department of Information Science Technology, University of Houston, Sugar Land, TX 77479, USA
[4] School of Computing, Clemson University, Clemson, SC 29634, USA
[5] Department of Public Health, Texas Tech University, Abilene, TX 79601, USA
[6] Department of Dermatology and Pathology, Texas Tech University, Abilene, TX 79601, USA
[7] Department of Dermatology, University of Texas MD Anderson Cancer Center, Houston, TX 77030, USA
[8] Department of Biomedical Engineering, University of Houston, Houston, TX 77204, USA
* Correspondence: zouridakis@uh.edu

Abstract

Melanoma is the most lethal form of skin cancer, and early detection is critical for improving patient outcomes. Although dermoscopy combined with deep learning has advanced automated skin-lesion analysis, progress is hindered by limited access to large, well-annotated datasets and by severe class imbalance, where melanoma images are substantially underrepresented. To address these challenges, we present the first systematic benchmarking study comparing four GAN architectures—DCGAN, StyleGAN2, and two StyleGAN3 variants (T and R)—for high-resolution (512 × 512) melanoma-specific synthesis. We train and optimize all models on two expert-annotated benchmarks (ISIC 2018 and ISIC 2020) under unified preprocessing and hyperparameter exploration, with particular attention to R1 regularization tuning. Image quality is assessed through a multi-faceted protocol combining distribution-level metrics (FID), sample-level representativeness (FMD), qualitative dermoscopic inspection, downstream classification with a frozen EfficientNet-based melanoma detector, and independent evaluation by two board-certified dermatologists. StyleGAN2 achieves the best balance of quantitative performance and perceptual quality, attaining FID scores of 24.8 (ISIC 2018) and 7.96 (ISIC 2020) at $\gamma = 0.8$. The frozen classifier recognizes 83% of StyleGAN2-generated images as melanoma, while dermatologists distinguish synthetic from real images at only 66.5% accuracy (chance = 50%), with low inter-rater agreement ($\kappa = 0.17$). In a controlled augmentation experiment, adding synthetic melanoma images to address class imbalance improved melanoma detection AUC from 0.925 to 0.945 on a held-out real-image test set. These findings demonstrate that StyleGAN2-generated melanoma images preserve diagnostically relevant features and can provide a measurable benefit for mitigating class imbalance in melanoma-focused machine learning pipelines.

Keywords: melanoma; skin cancer detection; synthetic data; generative adversarial networks; image synthesis; class imbalance

Check for updates

Academic Editor: Muneer Ahmad

Received: 23 December 2025
Revised: 11 February 2026
Accepted: 12 February 2026
Published: 20 February 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Melanoma accounts for only a small fraction of skin cancer diagnoses (about 6%) [1], yet it is responsible for the majority of skin-cancer-related deaths [1,2]. In the United States,

an estimated 186,680 melanoma cases were diagnosed in 2023, resulting in 7990 deaths [1]. Because the 5-year survival rate exceeds 99% when melanoma is detected early [1], timely and accurate recognition of suspicious lesions is clinically critical. However, melanoma remains difficult to detect reliably due to substantial intra-class variability (e.g., color, texture, borders) and frequent atypical presentations that overlap visually with benign lesions [3]. Automated screening systems based on dermoscopic criteria such as the 7-point checklist have shown promise for early detection [4–6], but their performance depends critically on the availability of large, diverse training datasets.

Recent advances in machine learning have improved automated dermoscopic image analysis, but progress is constrained by limited access to high-quality expert annotations and by severe class imbalance [7–9]: melanoma images are typically far fewer than non-melanoma cases in commonly used datasets. This imbalance can bias learned decision boundaries and reduce generalization, especially when models are deployed across acquisition settings, devices, and patient populations.

Generative modeling has therefore emerged as a promising direction to mitigate data scarcity and imbalance. In particular, Generative Adversarial Networks (GANs) can synthesize dermoscopic images that resemble real lesions and may enrich melanoma-specific variability for training and benchmarking [10,11]. Nevertheless, many prior approaches rely on relatively early or constrained generator designs that either produce limited-resolution images failing to preserve fine-grained dermoscopic cues, or depend heavily on feature-space metrics alone, which may not reflect clinical usefulness or downstream recognizability [12,13]. Moreover, melanoma-focused synthesis and systematic cross-architecture comparisons remain relatively underexplored.

In this study, we present the first systematic benchmark comparing four GAN-based architectures for high-resolution (512 $\times$ 512) melanoma image synthesis, while these architectures are well-established in the general computer vision literature, their relative performance for melanoma-specific synthesis—where preservation of fine-grained dermoscopic features is essential—has not been comprehensively evaluated. We train and optimize all models on two expert-annotated datasets (ISIC 2018 and ISIC 2020) under unified preprocessing and hyperparameter exploration, enabling direct cross-architecture comparison. Beyond standard generative metrics, we employ a multi-faceted evaluation protocol combining distribution-level assessment (FID), sample-level representativeness (FMD), qualitative dermoscopic inspection, downstream validation using a frozen EfficientNet-based melanoma classifier, and independent assessment by board-certified dermatologists. This combination of systematic comparison and clinically grounded evaluation addresses a gap in the literature, where prior studies often evaluate single architectures or rely solely on feature-space metrics.

The main contributions of this work are threefold: (1) a systematic cross-architecture comparison of DCGAN, StyleGAN2, and StyleGAN3 variants (T and R) for melanoma-specific image synthesis under consistent experimental conditions, with empirical evidence that StyleGAN2 provides the best balance of quantitative performance, perceptual quality, and artifact avoidance; (2) domain-specific hyperparameter optimization, particularly regarding R1 regularization strength (γ), with practical guidance for melanoma synthesis; and (3) a multi-faceted evaluation protocol combining distribution-level metrics, sample-level representativeness, downstream classifier validation, and independent assessment by two board-certified dermatologists. This demonstrated that 83% of StyleGAN2-generated images are recognized as melanoma by a strong external classifier and that expert dermatologists distinguish synthetic from real images at only 66.5% accuracy.

2. Related Work

2.1. GANs in Medical Imaging

The prevalence of severe class imbalance has motivated a shift from traditional geometric data augmentation toward generative modeling techniques. Early studies employed DCGANs to generate dermoscopic images for skin lesion classification but achieved limited realism due to low image resolution and insufficient preservation of fine-grained diagnostic structures [14].

More recently, Behara et al. [15] proposed an improved DCGAN classifier for skin lesion synthesis, demonstrating that careful hyperparameter tuning and image preprocessing can enhance DCGAN performance on dermatological datasets. Conditional GANs have also been applied to melanoma-specific tasks. For example, Ali et al. [16] utilized cGANs for melanoma lesion segmentation in IoMT-based systems, illustrating the versatility of adversarial frameworks across different medical imaging objectives.

Similar limitations were observed in other GAN-based dermatology studies, where low-resolution synthesis constrained clinical applicability [17]. To address resolution and stability issues, Progressive Growing GANs (PGGANs) were introduced into medical imaging. PGGANs enabled stable synthesis of high-resolution dermoscopic images, significantly improving visual fidelity and structural consistency compared to DCGANs [18,19]. Subsequently, StyleGAN-based models enabled fine-grained manipulation of lesion morphology and appearance [20]. StyleGAN-ADA demonstrated strong performance on limited medical datasets by dynamically adapting data augmentation strategies [21]. Beyond dermatology, FundusGAN has been applied to retinal imaging, preserving complex vascular structures and enabling effective augmentation for ophthalmic disease classification [22].

2.2. Emerging Alternatives: Diffusion Models

Denoising Diffusion Probabilistic Models (DDPMs) have recently emerged as a powerful alternative to GANs for image synthesis. Dhariwal and Nichol [23] demonstrated that diffusion models can surpass GANs on standard benchmarks such as ImageNet, achieving state-of-the-art FID scores through architectural improvements and classifier guidance. Latent diffusion models further improved computational efficiency by operating in compressed latent spaces [24], enabling high-resolution synthesis with reduced memory requirements. Akrout et al. [25] evaluated diffusion-based augmentation for skin disease classification, finding that synthetic images can match classifier performance when appropriately curated. Farooq et al. [26] proposed Derm-T2IM, a text-to-image framework using Stable Diffusion to generate melanoma and benign lesion images from natural language prompts. More recently, Wang et al. [27] applied diffusion-based augmentation specifically to address underrepresentation of minority subgroups in skin lesion datasets.

Despite these promising developments, diffusion models for dermatology remain in early stages relative to GAN-based approaches, with limited systematic evaluation on melanoma-specific synthesis tasks. Furthermore, diffusion models introduce different trade-offs: while they offer improved training stability, they typically require substantially longer inference times, and their ability to preserve fine-grained dermoscopic features has not been extensively validated. The present study therefore focuses on GAN architectures, which remain the most thoroughly characterized family for medical imaging synthesis, while acknowledging diffusion-based methods as a promising direction for future investigation.

2.3. Synthetic Data Generation for Melanoma Imaging

Severe data scarcity and class imbalance in melanoma imaging have motivated generative modeling as an alternative to purely geometric augmentation. However, early applications of classic GAN backbones often struggled to faithfully reproduce fine-grained

diagnostic structures, especially at higher resolutions [12]. To improve resolution and fidelity, progressive-growing strategies have been explored. Fumagal-Gonzalez et al. [13] employed PGGAN for melanoma synthesis and investigated how different real-to-synthetic ratios affect downstream melanoma detection, while PGGAN yielded visually richer samples, the authors observed that performance gains were not monotonic with increasing synthetic data and noted occasional suboptimal generations under practical training and hardware constraints, underscoring the remaining stability and consistency challenges of high-resolution GAN training. More recently, Abbasi et al. [28] fine-tuned a pre-trained Stable Diffusion model with LoRA and reported melanoma image generation with improved fine details, suggesting that large diffusion backbones can better capture complex lesion appearance. At the same time, diffusion models raise new practical considerations such as computational footprint and the need for careful domain validation when deployed for medical data augmentation. Alongside model development, Luschi et al. [29] proposed a holistic validation protocol for GAN-generated melanoma images that integrates objective computational metrics with structured expert assessment. Importantly, this line of work primarily advances how to validate synthetic melanoma images; it does not provide a controlled, large-scale benchmark of modern GAN architectures under consistent training protocols.

In summary, although generative models have advanced from DCGAN to PGGAN and, more recently, diffusion-based approaches, the literature still lacks systematic and controlled comparisons of state-of-the-art GANs for melanoma image synthesis. The present study addresses this gap.

3. Methods

3.1. Generative Models

Generative Adversarial Networks (GANs) [30] consist of two neural networks, a generator (G) and a discriminator (D), trained simultaneously in a competitive manner. The generator aims to produce synthetic data whose distribution closely resembles that of the real data, while the discriminator attempts to distinguish between real and generated samples. The training process can be formulated as a minimax game, where both G and D iteratively optimize their respective objectives. GAN training can be expressed mathematically as the following optimization problem [31]:

$$\min_G \max_D V(D,G) = \mathbb{E}_{x\sim p_{\text{data}}(x)}[\log D(x)] + \mathbb{E}_{z\sim p_z(z)}[\log(1-D(G(z)))]$$

In this formulation, the generator G aims to minimize the objective function by creating data that are indistinguishable from real samples, while the discriminator D seeks to maximize it by accurately distinguishing real data x from generated samples $G(z)$.

The term $\mathbb{E}_x[\log D(x)]$ quantifies the discriminator's success in recognizing real data, whereas $\mathbb{E}_z[\log(1-D(G(z)))]$ measures its ability to identify generated data as fake, where z denotes the input noise vector.

This adversarial process encourages the generator to improve the quality of synthesized data and the discriminator to enhance its classification accuracy, thereby improving overall GAN performance. The two networks are trained simultaneously through this iterative adversarial procedure, as illustrated in Figure 1.

The DCGAN model:Deep Convolutional Generative Adversarial Networks (DCGANs) employ convolutional neural networks in both the generator and discriminator [32] to capture spatial image structure, including edges, textures, and object relationships. Our implementation is shown in Figure 2 along with the specific parameters used.

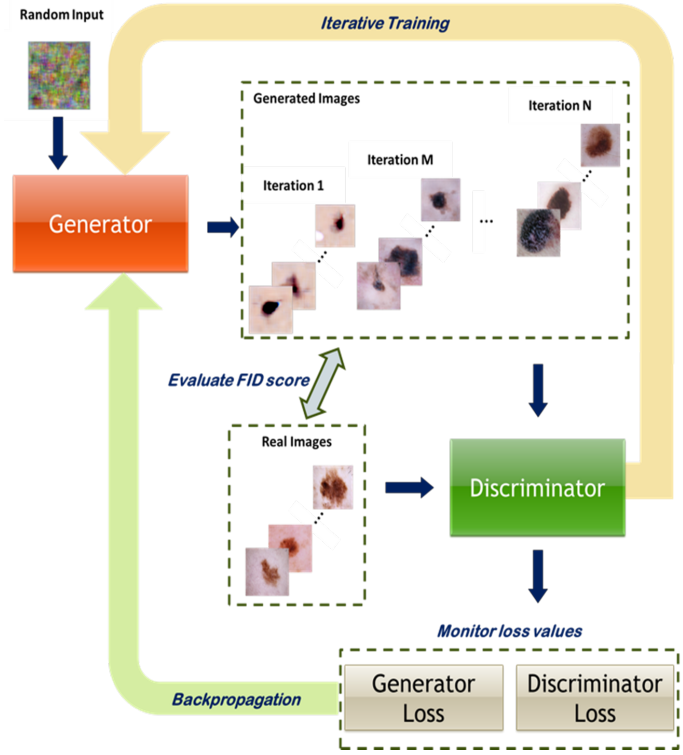


Figure 1. Iterative GAN training: the generator and discriminator undergo concurrent adversarial training.

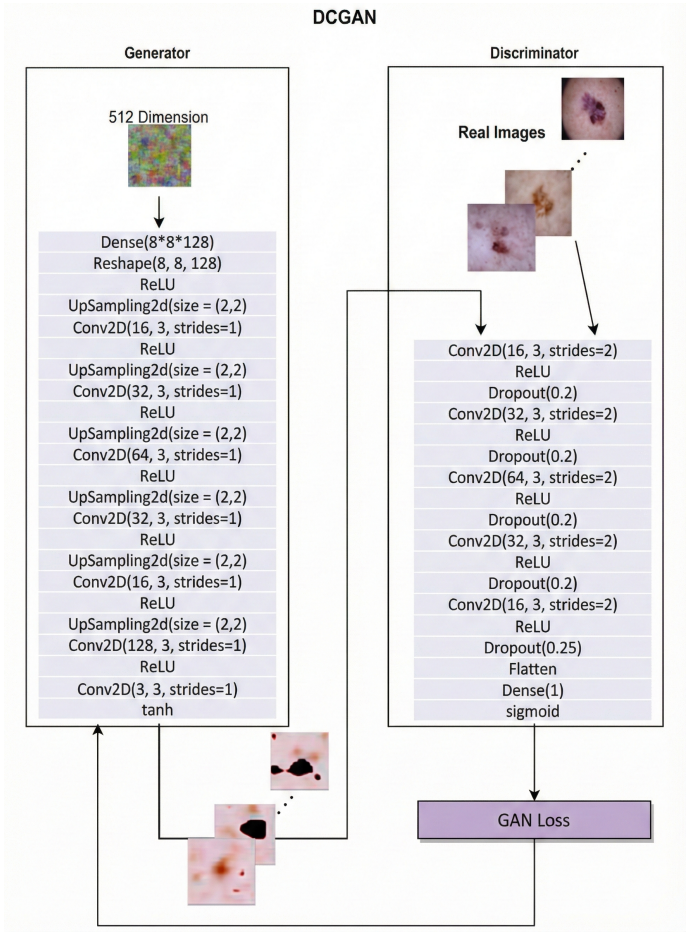


Figure 2. DCGAN architecture showing the generator and discriminator networks, including layer configuration, upsampling and strided convolutions, and output activations.

The StyleGAN models: The Style-based Generative Adversarial Network (StyleGAN) disentangles different aspects of image generation, such as content and style, through Adaptive Instance Normalization (AdaIN) layers that enable precise control over the output. The general architecture is shown in Figure 3. A random noise input vector (z) passes through the mapping network and is transformed into an intermediate latent space. The synthesis network (Figure 3b) progressively adds layers that increase the resolution of generated images (Figure 3c). The discriminator takes an image (real or generated) as input and produces a single output value representing the probability that the input image is real. The original StyleGAN uses a combined approach for content and style manipulation that limits independent adjustment; accordingly, it was not included in our evaluation. Instead, we selected StyleGAN2, which introduces a modular architecture that separates content and style [20,33]. Both models utilize progressive growing for high-resolution image generation, but StyleGAN2 employs enhanced upsampling techniques for sharper details. StyleGAN3 [34] is the latest version in the series and addresses the issue of "texture sticking," where repetitive patterns appear in some StyleGAN2-generated images. It relies on an alias-free generator architecture that uses Fourier features to represent image content.

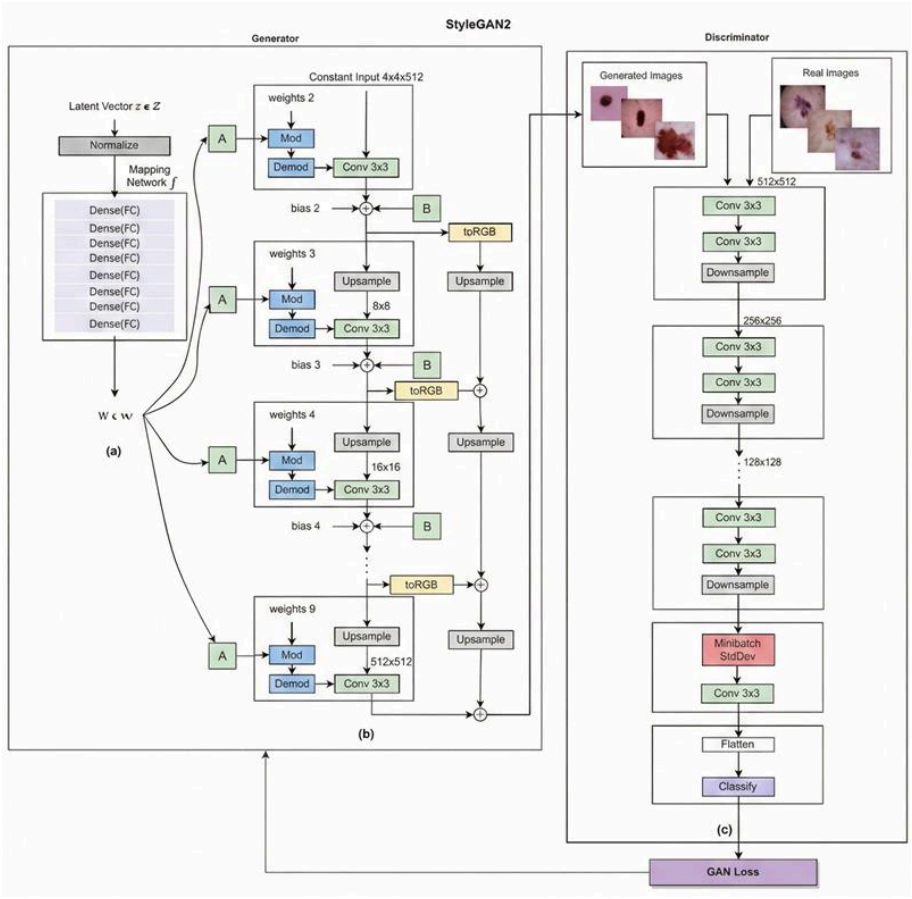


Figure 3. StyleGAN2 architecture illustrating the generator and discriminator networks. In the generator, (**a**) denotes the learned affine transformation used for style modulation, (**b**) represents the learned per-channel scaling factor, and ⊕ indicates element-wise addition. In the discriminator, (**c**) denotes the final classification stage.

3.2. Experimental Setup

3.2.1. Compute Cluster

Experiments were conducted on a DGX-2 system with 16 NVIDIA Tesla V100 GPUs (combined processing power of 2 petaflops), 512 GB of total GPU memory, 1.5 TB of NVMe storage, 15 TB of SATA storage, and the NVIDIA CUDA software, v12.2 stack (Nvidia

Corporation, Santa Clara, CA, USA). Four GPUs, each with 32 GB memory, were used for all model training.

3.2.2. Datasets

The ISIC 2018 dataset [35], provided by the International Skin Imaging Collaboration (ISIC), comprises 10,015 high-quality images representing seven different skin lesion types, including 1113 melanomas. The images are highly variable in terms of lighting, resolution, and lesion appearance. Images are annotated with a specific diagnosis and may include lesion localization, patient age, and sex. The second dataset, ISIC 2020 [36], contains 33,126 dermoscopic skin lesion images, including 7227 melanomas, all with associated metadata.

3.2.3. Data Preprocessing

For model training, we used 1061 melanoma images from the ISIC 2018 dataset after excluding 52 images with excessive artifacts, poor focus, or non-standard framing. All images were resized from their original dimensions to 512×512 pixels.

For DCGAN training, we employed bilateral filtering [37,38] to reduce noise while preserving edge structures, followed by image normalization across the dataset. The training set was augmented to 8488 images through rotations (90°, 180°, and 270°) and horizontal flipping. For StyleGAN2 and StyleGAN3 training, we applied only horizontal flipping to a random subset of images, as these architectures incorporate internal augmentation mechanisms that reduce the need for extensive external augmentation. This difference in augmentation strategy reflects architecture-specific best practices rather than an experimental variable; however, it should be considered when interpreting cross-architecture comparisons. Augmentation improved model robustness and reduced overfitting to specific orientations of melanoma patterns. Model parameters were further optimized using 7227 images from the ISIC 2020 dataset with the same preprocessing procedure.

3.2.4. Model Parameter Exploration

For the DCGAN model, we configured the kernel size to $(3, 3)$ and used LeakyReLU activation in the convolution layers, with tanh and sigmoid activations in the output layers of the generator and discriminator, respectively. The hyperparameters were set as follows: noise vector size of 512, batch size of 32, and maximum training iterations of 300,000—training images were repeatedly sampled from the melanoma dataset until this threshold was reached. We used the TruncatedNormal initializer, binary cross-entropy loss function, and the Adam optimizer with a learning rate of 0.0002 and $\beta_1 = 0.5$ [39]. We explored output resolutions ranging from 256×256 to 512×512 pixels, several dropout rates, different filter values for each layer, and the effects of batch normalization in both the generator and discriminator.

For the StyleGAN2 and StyleGAN3 models, we fixed the batch size at 32 and focused on optimizing the R1 regularization weight (γ), which plays a critical role in stabilizing training. We explored $\gamma \in \{0.8, 1.6, 8.0, 10.0\}$ based on recommendations in [34]. StyleGAN3 was trained using two configurations: StyleGAN3-R (rotation and translation equivariance), designed to minimize positional bias and improve rotational consistency [34]; and StyleGAN3-T (translation equivariance), which emphasizes realistic textures and fine details [34]. The maximum number of training images was set to 6,800,000, with images repeatedly sampled from the melanoma training dataset until this count was reached.

3.3. Model Evaluation

We used the Fréchet Inception Distance (FID) [40], which quantifies similarity between the distributions of real and synthetic images by comparing deep features extracted from the pre-trained Inception V3 network [41]. FID is computed as:

$$\text{FID} = \|\mu_r - \mu_g\|^2 + \text{Tr}\left(\Sigma_r + \Sigma_g - 2(\Sigma_r\Sigma_g)^{1/2}\right)$$

where (μ_r, Σ_r) and (μ_g, Σ_g) are the mean and covariance of the real and generated feature distributions, respectively. Lower FID indicates greater similarity to the real data distribution. However, FID scores are not directly interpretable in terms of human perception, and a lower score does not always guarantee that generated images will appear more realistic or prove useful in practical applications.

In addition to FID, we monitored generator and discriminator loss values across parameter settings to ensure loss convergence and identify potential model collapse, training instability, or discriminator dominance.

To address limitations of single-metric evaluation, recent work has emphasized the importance of multi-faceted assessment for synthetic medical images. Abdusalomov et al. [42] highlighted that existing metrics primarily evaluate distributional similarity but may fail to capture whether synthetic images preserve medically relevant features or introduce artifacts affecting downstream utility. Following these recommendations, we complement FID with the Fréchet Medoid Distance (FMD), which measures the distance from each generated sample to the medoid (most central sample) of the real distribution in feature space, providing a sample-level measure of representativeness that is more sensitive to mode collapse than distribution-level metrics. We further include qualitative dermoscopic inspection to identify clinical feature preservation, and downstream classifier evaluation to test retention of discriminative cues.

4. Results

4.1. FID and FMD Performance Analysis

To comprehensively evaluate the quality of generated images, we jointly consider the Fréchet Inception Distance (FID) and the Fréchet Medoid Distance (FMD), which quantify complementary aspects of generative performance. FID measures how closely the global feature distribution of generated samples matches that of real images by computing the Fréchet distance between Gaussian approximations in Inception feature space; lower values indicate better overall realism and diversity. In contrast, FMD evaluates sample-level representativeness by measuring distances to medoid real samples in feature space, making it more sensitive to mode collapse and local mismatches; again, lower values indicate better performance.

Table 1 reports the results for all models under $\gamma = 8$, which serves as a common reference point for cross-architecture comparison. In terms of FID, StyleGAN3-R achieves the lowest score (26.47), with StyleGAN2 close behind (31.58), both substantially outperforming DCGAN (66.49), while StyleGAN3-T performs markedly worse (246.42).

Interestingly, StyleGAN3-T presents a divergent pattern: while achieving the worst FID score (246.42), it obtains the lowest FMD (41.37). This apparent discrepancy reflects the complementary nature of these metrics. FID evaluates distributional similarity by comparing Gaussian approximations of feature distributions, penalizing models that fail to capture the full diversity of the real data. FMD, in contrast, measures sample-level representativeness by computing distances to medoid real samples, rewarding individual images that closely resemble typical real examples regardless of overall diversity. The combination of poor FID and strong FMD for StyleGAN3-T suggests limited mode coverage:

the model generates samples that individually resemble real melanomas but fails to capture the full morphological diversity present in the training distribution—a pattern consistent with partial mode collapse. StyleGAN2 and StyleGAN3-R achieve strong performance on both metrics (FID: 31.58 and 26.47; FMD: 50.08 and 49.21, respectively), indicating both distributional fidelity and sample-level representativeness. DCGAN performs poorly on both metrics, reflecting limited capacity for high-resolution medical image synthesis.

Although StyleGAN3-R achieves the lowest FID at $\gamma = 8$, qualitative inspection of generated samples (Figure 4) reveals prominent mesh- or grid-like artifacts. These patterns violate fundamental realism requirements for dermoscopic images and render the outputs unsuitable for medical applications, yet they are not adequately penalized by either feature-space metric. Visual inspection of 100 randomly selected StyleGAN3-R samples revealed such artifacts in approximately 60% of images. For this reason, we select StyleGAN2 as the most reliable model overall, balancing strong quantitative performance with stable perceptual quality.

Having identified StyleGAN2 as the preferred architecture, we investigated the effect of R1 regularization strength (γ) on its performance. Figure 5 shows the evolution of FID with respect to training set size, demonstrating that StyleGAN2 benefits consistently from additional data, indicating robust scaling behavior. Furthermore, Table 2 demonstrates that smaller γ values yield better FID scores on both ISIC 2018 and ISIC 2020, with $\gamma = 0.8$ producing the best results (24.8 and 7.96, respectively). This finding suggests that lighter regularization is preferable for melanoma synthesis, likely because the relatively homogeneous dermoscopic domain requires less aggressive smoothing of the discriminator's gradients.

Table 1. Performance comparison across architectures at $\gamma = 8$. Lower FID and FMD values correspond to better performance (↓).

Metric	DCGAN	StyleGAN2	StyleGAN3-T	StyleGAN3-R
FID ↓	66.49	31.58	246.42	26.47
FMD ↓	695.93	50.08	41.37	49.21

Table 2. FID scores for StyleGAN2 under different R1 regularization strengths (γ) on the ISIC 2018 and ISIC 2020 datasets. Lower γ values yield lower FID scores.

γ	FID (ISIC 2018)	FID (ISIC 2020)
0.8	24.8	7.96
1.6	27.4	9.48
8.0	31.6	9.91
10.0	33.2	10.4

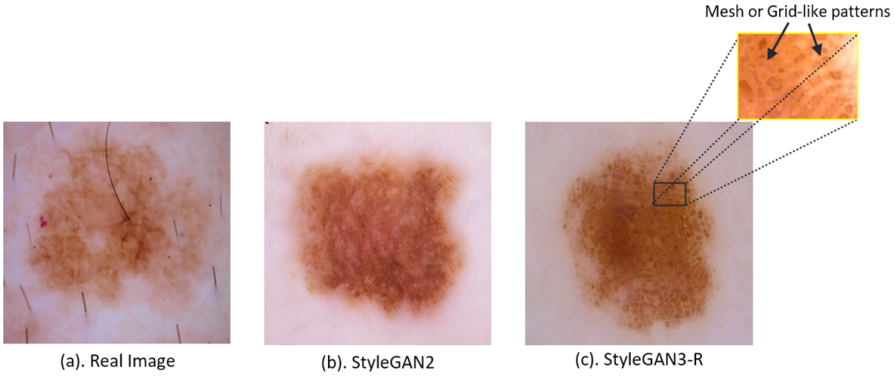


Figure 4. (**a**) Real melanoma and images produced by (**b**) StyleGAN2 and (**c**) StyleGAN3-R. The zoomed inset highlights mesh- or grid-like artifacts present in StyleGAN3-R outputs.

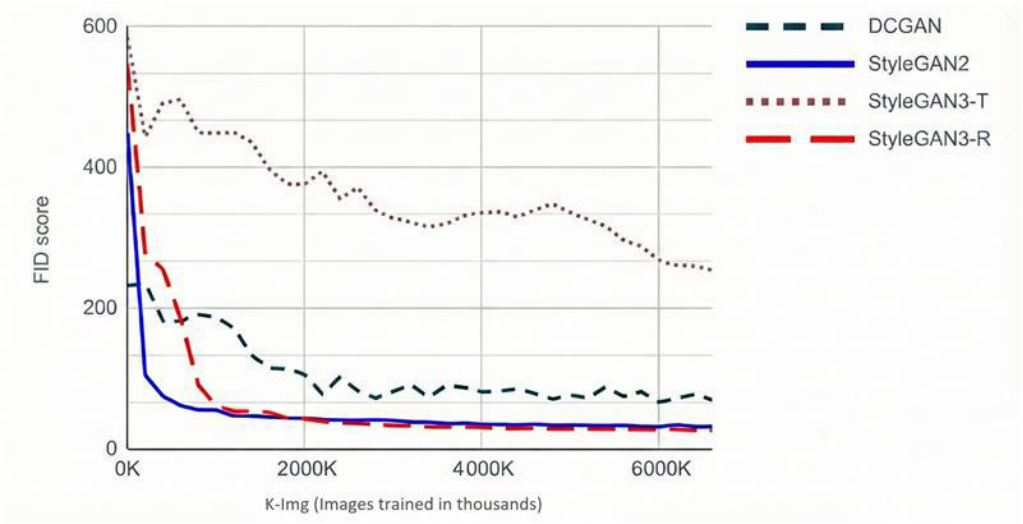


Figure 5. FID score as a function of training images (thousands) for each architecture. StyleGAN2 shows consistent improvement with additional data, while StyleGAN3-T plateaus at substantially higher FID values.

4.2. Image Generation

Each model produced 1000 synthetic melanoma images per parameter configuration. Figure 6 shows representative examples from each model alongside real melanomas used for training. We evaluated the quality of synthetic images by examining the presence of characteristic dermoscopic features captured by the 7-point checklist [43], which dermatologists use for melanoma diagnosis.

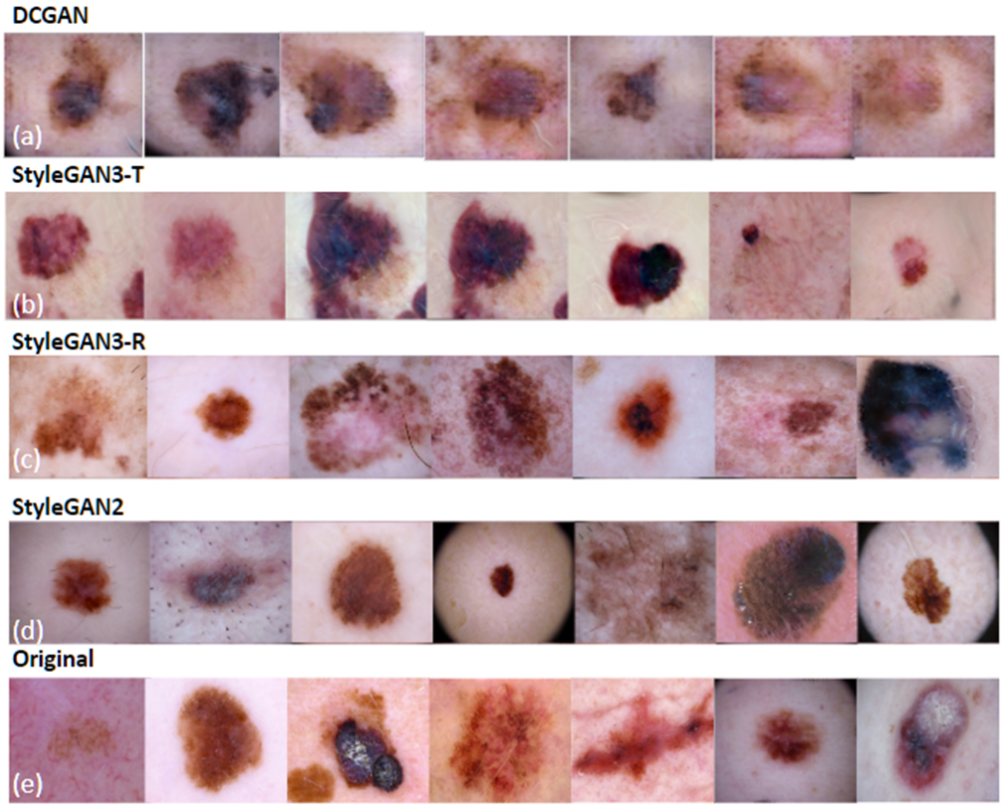


Figure 6. Representative synthetic images from (**a**) DCGAN, (**b**) StyleGAN3-T, (**c**) StyleGAN3-R, and (**d**) StyleGAN2, compared with (**e**) real melanoma images used for training.

Despite training on over 6 million image presentations, DCGAN and StyleGAN3-T outputs lack the fine details expected in melanoma images, consistent with their elevated FID scores. In contrast, images produced by StyleGAN2 and StyleGAN3-R consistently exhibit the high-quality dermoscopic details present in real melanomas, including pigment network patterns, color variegation, and border irregularity. However, as noted above, StyleGAN3-R generates a substantial proportion of images with mesh- or grid-like artifacts (Figure 4), attributable to the model's strict enforcement of translation and rotation equivariance constraints.

Overall, among all four architectures evaluated, StyleGAN2 produced the most realistic melanoma images in terms of the major and minor dermoscopic features of the 7-point checklist—an assessment corroborated by its strong FID scores.

4.3. Computational Cost and Parameter Size

We evaluated the computational efficiency of different GAN architectures in terms of training time and model size (Table 3). Using the ISIC 2018 dataset, DCGAN required 0.9 h to train, StyleGAN2 required 2.8 h to reach its optimal FID score, and both StyleGAN3-R and StyleGAN3-T required 9.2 h each.

The parameter count is reported as $G + D$, where G and D denote the number of parameters (in millions) in the generator and discriminator, respectively. DCGAN has a substantially smaller model size (5M total), while StyleGAN-based models employ larger networks (54–59M total), resulting in higher computational cost. Notably, StyleGAN2 achieves the best quality–efficiency tradeoff: it requires only 30% of the training time of StyleGAN3 variants while producing superior or comparable output quality.

Table 3. Model size and training time for each GAN architecture on ISIC 2018.

	DCGAN	StyleGAN2	StyleGAN3-T	StyleGAN3-R
Parameters (M), $G + D$	4 + 1	30 + 29	25 + 29	25 + 29
Training time (hours)	0.9	2.8	9.2	9.2

5. Downstream Evaluation

5.1. Evaluation Model: External Skin Lesion Classifier

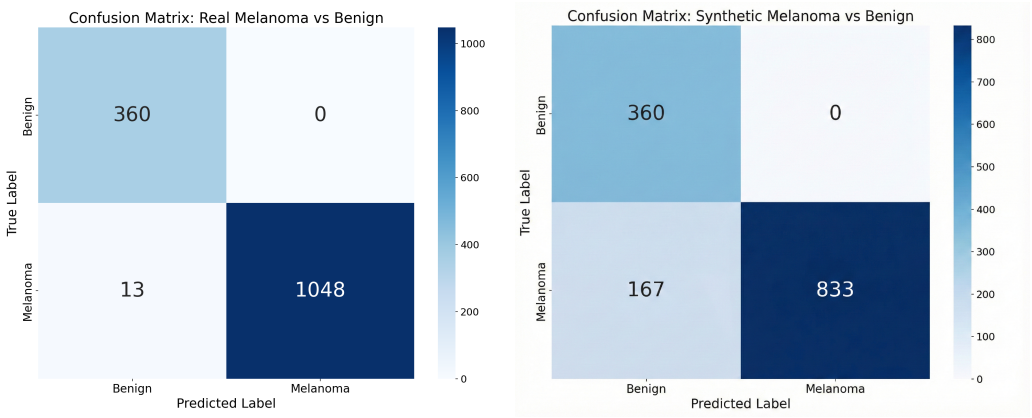
To test whether synthetic melanoma images retain discriminative characteristics beyond feature-space similarity scores, we employed a strong external skin-lesion classifier as a downstream evaluator [44]. This model is an EfficientNet-B6-based ensemble developed for the SIIM-ISIC Melanoma Classification Challenge, where it achieved an AUC of 0.9490 on the private leaderboard, ranking among the top solutions. This strong baseline performance on real dermoscopic data makes it a rigorous test of whether synthetic images preserve melanoma-discriminative features. For consistency with our experimental setting, we map the model's outputs into two classes: melanoma and benign.

5.2. Downstream Evaluation I: Recognizability Under a Frozen Classifier

We first evaluated the classifier's decision behavior using its pretrained weights under two test configurations: (i) a *Real set* containing real benign ($n = 360$) and real melanoma ($n = 1061$) images, and (ii) a *Synthetic set* where the benign subset is identical but the melanoma images are replaced by 1000 randomly sampled StyleGAN2-generated images. Figure 7 summarizes the results.

On the Real set, the classifier achieves near-ceiling performance: all 360 benign images are correctly classified, while 98.8% (1048/1061) of real melanomas are correctly identified, establishing a strong baseline within our test distribution. On the Synthetic set, melanoma sensitivity decreases to 83.3% (833/1000), with 16.7% of synthetic melanomas misclassified

as benign. Nevertheless, the majority of generated samples are recognized as melanoma by this strong real-trained model.



(**a**) (**b**)

Figure 7. Confusion matrices for the frozen external classifier on (**a**) the Real set (real melanoma vs. real benign) and (**b**) the Synthetic set (StyleGAN2-generated melanoma vs. real benign). The classifier recognizes 83% of synthetic melanomas as melanoma.

This frozen-classifier evaluation directly measures whether synthetic melanoma images fall within the melanoma-relevant decision regions learned from real dermoscopic data, providing task-level evidence that the generated images preserve discriminative disease cues. This complements feature-space metrics (FID, FMD), which assess distributional similarity but not necessarily diagnostic relevance. Importantly, recognizability under a fixed classifier is a necessary—though not sufficient—condition for augmentation utility: if synthetic images are not recognized as melanoma by a strong real-trained model, they are unlikely to improve downstream training when used for data augmentation.

5.3. Downstream Evaluation II: Augmentation Utility

Motivated by the recognizability results, we further evaluated whether StyleGAN2-generated melanoma images can improve classifier performance when used as training augmentation. We trained the EfficientNet-B0 classifier at 512×512 resolution under two controlled training regimes:

- Real-only: The model is trained exclusively on real images ($n \approx 20{,}000$), resulting in a highly imbalanced class distribution with a benign-to-melanoma ratio of approximately 98:2.
- Real + Synthetic: The training set combines all real images with 6500 StyleGAN2-generated synthetic melanoma images, yielding a more balanced benign-to-melanoma ratio of approximately 65:35.

Both training sets were drawn from ISIC 2018 and ISIC 2020, with a random 80/10/10 split for training, validation, and testing. The test set ($n = 2000$; 1960 benign, 40 melanoma) consisted exclusively of real images and was held out from all GAN training to ensure unbiased evaluation. Both classifiers were trained for 20 epochs using identical hyperparameters (Adam optimizer, learning rate 10^{-4}, batch size 32), with the best checkpoint selected based on validation AUC.

Table 4 reports the results. The model trained with Real + Synthetic data achieved 98.27% overall accuracy and a melanoma AUC of 0.9445, compared to 85.07% accuracy and AUC of 0.9252 for the Real-only model. The F1 score for melanoma detection improved from 0.1682 to 0.2586.

Table 4. Classifier performance on the held-out real-image test set. The test set contains 1960 benign and 40 melanoma images, reflecting real-world class imbalance.

Training Data	Accuracy (%) †	Melanoma AUC	Melanoma F1
Real-only	85.07	0.9252	0.1682
Real + Synthetic	98.27	0.9445	0.2586

† Accuracy is dominated by the benign majority class; AUC is more informative for imbalanced data.

Several considerations apply when interpreting these results. First, overall accuracy is dominated by the benign majority class (98% of test samples), making it a poor measure of melanoma detection ability; melanoma AUC is more informative as it measures the model's ability to rank melanomas above benign samples across all decision thresholds. Second, the relatively low F1 scores (even after augmentation) reflect the extreme class imbalance in the test set: with only 40 melanoma samples, even a small number of false negatives or false positives substantially impacts precision and recall. Third, the improvement in AUC from 0.9252 to 0.9445 represents a meaningful gain, though formal statistical testing (e.g., DeLong's test) would require a larger melanoma test set for adequate power.

Taken together, these results support the claim that StyleGAN2-generated melanoma images are not only recognizable by a strong evaluator but can also provide measurable downstream utility when used to address class imbalance in a controlled training setup.

6. Dermatologist Evaluation

To assess the perceptual realism of GAN-generated melanoma images, we constructed a balanced evaluation set of 200 images consisting of 100 real melanomas (randomly sampled from ISIC 2018) and 100 synthetic melanomas (randomly sampled from StyleGAN2 outputs at $\gamma = 0.8$). Images were presented in randomized order without any identifying information. We report classification accuracy for (i) a machine baseline and (ii) two board-certified dermatologists who performed the task independently.

6.1. Machine Baseline: StyleGAN2 Discriminator

As an initial reference point, we evaluated the trained StyleGAN2 discriminator on the real-versus-synthetic classification task. The discriminator achieved an overall accuracy of 59.5% (Table 5), only modestly above chance (50%). Notably, the discriminator exhibited an asymmetric error pattern: it achieved 84.0% accuracy on synthetic images but only 35.0% on real images, indicating a bias toward classifying images as synthetic. This suggests that the generated images are sufficiently close to the training distribution that even the model's internal real/fake signal provides limited separability. We report the discriminator results *not* as an additional rater, but as a computational benchmark for contextualizing human performance under the same decision setting.

Table 5. Performance comparison of human raters and the StyleGAN2 discriminator on the real-versus-synthetic classification task ($n = 200$ images).

Metric	Dermatologist 1	Dermatologist 2	Human Mean	Discriminator
Overall Accuracy	71.0% ($p < 0.001$)	62.0% ($p < 0.001$)	66.5%	59.5%
Real Accuracy	51.0%	70.0%	60.5%	35.0%
Synthetic Accuracy	91.0%	54.0%	72.5%	84.0%
Accepted as Real †	9.0%	46.0%	27.5%	16.0%

† Percentage of synthetic images classified as real.

6.2. Independent Dermatologist Assessment

Two board-certified dermatologists independently labeled the 200-image set. Both raters are co-authors of this study and are affiliated with separate major academic medical centers, ensuring independent clinical perspectives. Neither rater had prior exposure to the synthetic images or knowledge of the real/synthetic ratio.

Binomial testing confirmed that both raters performed significantly above the 50% chance level (Dermatologist 1: 71.0%, $p < 0.001$; Dermatologist 2: 62.0%, $p < 0.001$; mean overall accuracy, 66.5%), indicating that their selections were deliberate rather than random, despite the difficulty in distinguishing synthetic from real melanoma images.

Notably, the two raters exhibited complementary decision tendencies: Dermatologist 1 achieved high accuracy on synthetic images (91.0%) but near-chance accuracy on real images (51.0%), consistent with a conservative strategy that preferentially flags images as synthetic. Dermatologist 2 showed the opposite pattern, with higher accuracy on real images (70.0%) but lower accuracy on synthetic images (54.0%), reflecting a more liberal threshold. These complementary labeling patterns indicate that classification difficulty is not confined to a single class and that consistent "tell-tale" artifacts are not present across synthetic samples.

The percentage of synthetic images accepted as real by each rater was 9.0% for Dermatologist 1 and 46.0% for Dermatologist 2, with a mean of 27.5%. This variability further underscores the absence of consistent visual markers distinguishing synthetic from real images.

6.3. Inter-Rater Reliability

To quantify agreement beyond chance, we computed Cohen's κ, defined as [45]:

$$\kappa = \frac{P_o - P_e}{1 - P_e},$$

where P_o is the observed agreement and P_e is the expected agreement under chance, given the raters' marginal label distributions. The resulting κ values are shown in Table 6. Statistical significance of κ was assessed using a Z-test, where $Z = \kappa / SE$ and the standard error is given by:

$$SE = \sqrt{\frac{P_o(1 - P_o)}{n(1 - P_e)^2}}.$$

Table 6. Inter-rater agreement (Cohen's κ) for the real-versus-synthetic classification task.

Comparison	Cohen's κ	p-Value	Agreement
Dermatologist 1 vs. Dermatologist 2	0.173	0.009	Slight
Dermatologist 1 vs. Discriminator	0.042	0.482	Slight
Dermatologist 2 vs. Discriminator	0.082	0.187	Slight

Inter-rater agreement between the two dermatologists was low but statistically significant ($\kappa = 0.173$, $p = 0.009$), indicating substantial variability in labeling criteria even among experts [46]. Agreement between each dermatologist and the discriminator was negligible and not statistically significant ($\kappa \leq 0.082$, $p > 0.05$), consistent with humans and the discriminator relying on different visual cues.

6.4. Summary

Both machine and human evaluations converge on the same conclusion: distinguishing StyleGAN2-generated melanoma images from real melanomas is difficult under visual inspection. The modest above-chance accuracy (66.5% human mean), low inter-rater

agreement ($\kappa = 0.173$), and complementary response patterns across raters collectively support the perceptual realism of the generated samples.

7. Limitations and Future Work

Certain tradeoffs of the analyzed generators should be noted. DCGAN exhibits limited capacity for high-resolution melanoma synthesis, often failing to preserve fine-grained dermoscopic details. StyleGAN3-T shows limited mode coverage, producing individually realistic samples while failing to capture the full diversity of melanoma appearances. StyleGAN3-R improves distributional fidelity but introduces mesh- or grid-like artifacts that are undesirable in medical images. StyleGAN2 achieves strong overall performance but remains sensitive to regularization settings.

Future work will extend this study in several directions. First, evaluating cross-dataset generalization to other imaging modalities (e.g., smartphone-captured images) and external datasets will assess the robustness of synthetic augmentation strategies. Second, given recent advances in diffusion-based generative models, a comparative evaluation of latent diffusion models against the GAN architectures benchmarked here will determine whether these newer approaches offer advantages for preserving fine-grained dermoscopic features. Third, extending to conditional generation would address specific gaps in training data, such as skin type, melanoma subtype, and anatomical location.

Finally, structured expert assessment using the 7-point dermoscopic checklist will validate clinical feature preservation and identify artifacts not captured by automated metrics. Additionally, integrating synthetic images with automated 7-point checklist detection systems [4,6,47] could determine whether GAN-generated melanomas preserve the specific dermoscopic features (e.g., atypical pigment network, asymmetry, vascular patterns) required for algorithmic classification.

8. Conclusions

This study presents the first systematic benchmark of four GAN architectures for high-resolution (512×512) melanoma image synthesis, addressing a critical bottleneck in dermatological AI: the scarcity of annotated melanoma images and severe class imbalance in training datasets. Using consistent protocols and multi-faceted evaluation on two expert-annotated benchmarks (ISIC 2018 [35] and ISIC 2020 [48]), we demonstrate that StyleGAN2 achieves the optimal balance of distributional fidelity, perceptual realism, and artifact avoidance, attaining FID scores of 24.8 and 7.96 on the respective datasets.

Three lines of evidence support the diagnostic relevance of StyleGAN2-generated melanomas: (1) a frozen EfficientNet-based classifier recognized 83% of synthetic samples as melanoma, confirming preservation of disease-discriminative features; (2) board-certified dermatologists from independent institutions distinguished synthetic from real images at only 66.5% accuracy, demonstrating the absence of consistent visual artifacts; and (3) augmenting a class-imbalanced training set with synthetic melanomas improved detection AUC from 0.925 to 0.945, providing direct evidence of downstream clinical utility.

These results demonstrate that high-quality synthetic melanoma images can serve as a practical tool for mitigating class imbalance in melanoma detection pipelines. As melanoma remains the deadliest form of skin cancer, with outcomes highly dependent on early detection, methods that improve automated screening systems have significant potential clinical impact. Moreover, the proposed framework may extend to other data-scarce dermatological conditions, such as Buruli ulcer disease, where the limited availability of annotated images similarly hinders the development of reliable automated screening tools [49]. This work establishes a foundation for integrating synthetic data into dermato-

logical AI development—not as a substitute for real patient data, but as a complementary resource for improving model robustness and generalization.

Author Contributions: Conceptualization, G.Z.; methodology, P.-Y.L., R.H. and S.H.; software, P.-Y.L., Y.S., N.M. and Y.S.; algorithm validation, Y.S.; clinical validation, C.E.W., A.C.; formal analysis, N.M. and R.H.; investigation, C.M.Q. and G.Z.; resources, Y.S. and N.M.; data curation, P.-Y.L. and N.M.; writing—original draft, P.-Y.L., G.Z.; writing—review and editing, Y.S., N.M., R.H., and G.Z.; supervision, R.H. and G.Z.; project administration, G.Z.; funding acquisition, C.M.Q. and G.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This work was partially supported by the Texas Tech University Innovation Hub, Prototype Fund, 2022.

Data Availability Statement: The datasets used in this study are publicly available from the ISIC archive [35,48].

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Siegel, R.L.; Kratzer, T.B.; Wagle, N.S.; Sung, H.; Jemal, A. Cancer statistics, 2026. *CA Cancer J. Clin.* **2026**, *76*, e70043. [CrossRef]
2. Siegel, R.L.; Miller, K.D.; Fuchs, H.E.; Jemal, A. Cancer Statistics, 2022. *CA Cancer J. Clin.* **2022**, *72*, 7–33. [CrossRef]
3. Luke, J.J.; Flaherty, K.T.; Ribas, A.; Long, G.V. Targeted agents and immunotherapies: Optimizing outcomes in melanoma. *Nat. Rev. Clin. Oncol.* **2017**, *14*, 463–482. [CrossRef]
4. Wadhawan, T.; Situ, N.; Rui, H.; Lancaster, K.; Yuan, X.; Zouridakis, G. Implementation of the 7-point checklist for melanoma detection on smart handheld devices. In *Proceedings of the 2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*; IEEE: New York, NY, USA, 2011; pp. 3180–3183.
5. Situ, N.; Wadhawan, T.; Yuan, X.; Zouridakis, G. Modeling spatial relation in skin lesion images by the graph walk kernel. In *Proceedings of the 2010 Annual International Conference of the IEEE Engineering in Medicine and Biology*; IEEE: New York, NY, USA, 2010; pp. 6130–6133.
6. Zouridakis, G. Methods for Screening and Diagnosing a Skin Condition. US Patent 10,593,040, 17 March 2020.
7. Zareen, S.S.; Hossain, M.S.; Wang, J.; Kang, Y. Recent innovations in machine learning for skin cancer lesion analysis and classification: A comprehensive analysis of computer-aided diagnosis. *Precis. Med. Sci.* **2025**, *14*, 15–40. [CrossRef]
8. Yao, P.; Shen, S.; Xu, M.; Liu, P.; Zhang, F.; Xing, J.; Shao, P.; Kaffenberger, B.; Xu, R.X. Single Model Deep Learning on Imbalanced Small Datasets for Skin Lesion Classification. *IEEE Trans. Med. Imaging* **2022**, *41*, 1242–1254. [CrossRef]
9. Raza, W.H.; Shah, A.B.; Wen, Y.; Shen, Y.; Lemus, J.D.M.; Schiess, M.C.; Ellmore, T.M.; Hu, R.; Fu, X. NeuroMoE: A Transformer-Based Mixture-of-Experts Framework for Multi-Modal Neurological Disorder Classification. *arXiv* **2025**, arXiv:2506.14970.
10. Innani, S.; Dutande, P.; Baid, U.; Pokuri, V.; Bakas, S.; Talbar, S.; Baheti, B.; Guntuku, S.C. Generative adversarial networks based skin lesion segmentation. *Sci. Rep.* **2023**, *13*, 13467. [CrossRef] [PubMed]
11. La Salvia, M.; Torti, E.; Leon, R.; Fabelo, H.; Ortega, S.; Martinez-Vega, B.; Callico, G.M.; Leporati, F. Deep Convolutional Generative Adversarial Networks to Enhance Artificial Intelligence in Healthcare: A Skin Cancer Application. *Sensors* **2022**, *22*, 6145. [CrossRef] [PubMed]
12. Rodriguez Alonso, A.; Sanchez Diez, A.; Cancho Galán, G.; Ibarrola Altuna, R.; Irigoyen Miró, G.; Penas Lago, C.; Boyano López, M.D.; Izu Belloso, R. Enhancing Melanoma Diagnosis in Histopathology with Deep Learning and Synthetic Data Augmentation. *Bioengineering* **2025**, *12*, 1001. [CrossRef] [PubMed]
13. Fumagal-González, G.A.; Garza-Abdala, J.A.; Pérez, A.A.P.; Tamez-Pena, J.G. Assessing a Proposed Dynamic Ratio in Dataset Class Imbalance with GAN's Generated Melanoma Images. In Proceedings of the 2025 IEEE 38th International Symposium on Computer-Based Medical Systems (CBMS), Madrid, Spain, 18–20 June 2025; pp. 500–503. [CrossRef]
14. Bissoto, A.; Perez, F.; Valle, E.; Avila, S. Skin Lesion Synthesis with Generative Adversarial Networks. In Proceedings of the IEEE International Symposium on Biomedical Imaging (ISBI), Venice, Italy, 8–11 April 2019; pp. 1–4.
15. Behara, K.; Bhero, E.; Agee, J.T. Skin Lesion Synthesis and Classification Using an Improved DCGAN Classifier. *Diagnostics* **2023**, *13*, 2635. [CrossRef]
16. Ali, Z.; Naz, S.; Zaffar, H.; Choi, J.; Kim, Y. An IoMT-Based Melanoma Lesion Segmentation Using Conditional Generative Adversarial Networks. *Sensors* **2023**, *23*, 3548. [CrossRef] [PubMed]
17. Frid-Adar, M.; Diamant, I.; Klang, E.; Amitai, M.; Goldberger, J.; Greenspan, H. GAN-Based Synthetic Medical Image Augmentation for Increased CNN Performance in Liver Lesion Classification. *Neurocomputing* **2018**, *321*, 321–331. [CrossRef]

18. Karras, T.; Aila, T.; Laine, S.; Lehtinen, J. Progressive Growing of GANs for Improved Quality, Stability, and Variation. In Proceedings of the International Conference on Learning Representations (ICLR), Vancouver, BC, Canada, 30 April–3 May 2018.
19. Bissoto, A.; Fornaciali, M.; Valle, E.; Avila, S. (De)constructing Bias on Skin Lesion Datasets. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Long Beach, CA, USA, 16–17 June 2019.
20. Karras, T.; Laine, S.; Aila, T. A Style-Based Generator Architecture for Generative Adversarial Networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–17 June 2019; pp. 4401–4410.
21. Karras, T.; Aittala, M.; Laine, S.; Härkönen, E.; Hellsten, J.; Lehtinen, J.; Aila, T. Training Generative Adversarial Networks with Limited Data. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), Virtual, 6–12 December 2020.
22. Costa, P.; Galdran, A.; Smailagic, A.; Campilho, A. A Weakly-Supervised Framework for Interpretability of Retinal Disease Classification. *IEEE Trans. Med. Imaging* **2018**, *37*, 2539–2551.
23. Dhariwal, P.; Nichol, A. Diffusion Models Beat GANs on Image Synthesis. In Proceedings of the Advances in Neural Information Processing Systems, Online, 6–14 December 2021; Volume 34, pp. 8780–8794.
24. Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; Ommer, B. High-Resolution Image Synthesis with Latent Diffusion Models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 19–20 June 2022; pp. 10684–10695.
25. Akrout, M.; Gyepesi, B.; Holló, P.; Poór, A.; Kincső, B.; Solis, S.; Cirone, K.; Kawahara, J.; Slade, D.; Abid, L.; et al. Diffusion-Based Data Augmentation for Skin Disease Classification: Impact Across Original Medical Datasets to Fully Synthetic Images. In *Proceedings of the Deep Generative Models, DGM4MICCAI 2023*; Lecture Notes in Computer Science; Springer: Berlin/Heidelberg, Germany, 2024; Volume 14533, pp. 99–109. [CrossRef]
26. Farooq, M.A.; Yao, W.; Schukat, M.; Little, M.A.; Corcoran, P. Derm-T2IM: Harnessing Synthetic Skin Lesion Data via Stable Diffusion Models for Enhanced Skin Disease Classification. *arXiv* **2024**, arXiv:2401.05159. [CrossRef]
27. Wang, J.; Chung, Y.; Ding, Z.; Hamm, J. From Majority to Minority: A Diffusion-based Augmentation for Underrepresented Groups in Skin Lesion Analysis. In *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention*; Springer Nature: Cham, Switzerland, 2024; pp. 14–23.
28. Abbasi, S.F.; Bilal, M.; Mukherjee, T.; Islam, S.u.; Pournik, O.; Arvanitis, T.N. Synthetic Image Generation for Skin Lesion Analysis Using Stable Diffusion Models. In *Proceedings of the Studies in Health Technology and Informatics*; IOS Press: Amsterdam, The Netherlands, 2025; Volume 323, pp. 81–85. [CrossRef]
29. Luschi, A.; Tognetti, L.; Cartocci, A.; Cinotti, E.; Rubegni, G.; Calabrese, L.; D'onghia, M.; Dragotto, M.; Moscarella, E.; Brancaccio, G.; et al. Design and development of a systematic validation protocol for synthetic melanoma images for responsible use in medical artificial intelligence. *Biocybern. Biomed. Eng.* **2025**, *45*, 608–616. [CrossRef]
30. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative Adversarial Nets. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), Montreal, QC, Canada, 8–13 December 2014; pp. 2672–2680.
31. Bastien, F.; Lamblin, P.; Pascanu, R.; Bergstra, J.; Goodfellow, I.; Bergeron, A.; Bouchard, N.; Warde-Farley, D.; Bengio, Y. Theano: New Features and Speed Improvements. *arXiv* **2012**, arXiv:1211.5590. [CrossRef]
32. Radford, A.; Metz, L.; Chintala, S. Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. *arXiv* **2015**, arXiv:1511.06434. [CrossRef]
33. Karras, T.; Laine, S.; Aittala, M.; Hellsten, J.; Lehtinen, J.; Aila, T. Analyzing and Improving the Image Quality of StyleGAN. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 14–19 June 2020; pp. 8110–8119.
34. Karras, T.; Aittala, M.; Laine, S.; Härkönen, E.; Hellsten, J.; Lehtinen, J.; Aila, T. Alias-Free Generative Adversarial Networks. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), Virtual, 6–14 December 2021.
35. Codella, N.; Rotemberg, V.; Tschandl, P.; Celebi, M.E.; Dusza, S.; Gutman, D.; Helba, B.; Kalloo, A.; Liopyris, K.; Marchetti, M.; et al. Skin Lesion Analysis Toward Melanoma Detection 2018: A Challenge Hosted by the International Skin Imaging Collaboration (ISIC). *arXiv* **2019**, arXiv:1902.03368. [CrossRef]
36. International Skin Imaging Collaboration (ISIC), SIIM-ISIC 2020 Challenge Dataset, 2020. Available online: https://www.isic-archive.com (accessed on 11 February 2025).
37. Tomasi, C.; Manduchi, R. Bilateral Filtering for Gray and Color Images. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Bombay, India, 4–7 January 1998; pp. 839–846.
38. Gavaskar, R.G.; Chaudhury, K.N. Fast Adaptive Bilateral Filtering. *IEEE Trans. Image Process.* **2018**, *28*, 779–790. [CrossRef]
39. Mutepfe, F.; Kalejahi, B.K.; Meshgini, S.; Danishvar, S. Generative Adversarial Network Image Synthesis Method for Skin Lesion Generation and Classification. *J. Med. Signals Sens.* **2021**, *11*, 237–252. [CrossRef]

40. Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; Hochreiter, S. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), Long Beach, CA, USA, 4–9 December 2017; pp. 6626–6637.
41. Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; Wojna, Z. Rethinking the Inception Architecture for Computer Vision. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 2818–2826.
42. Abdusalomov, A.B.; Nasimov, R.; Nasimova, N.; Muminov, B.; Whangbo, T.K. Evaluating Synthetic Medical Images Using Artificial Intelligence with the GAN Algorithm. *Sensors* **2023**, *23*, 3440. [CrossRef]
43. Argenziano, G.; Fabbrocini, G.; Carli, P.; De Giorgi, V.; Sammarco, E.; Delfino, M. Epiluminescence Microscopy for the Diagnosis of Doubtful Melanocytic Skin Lesions: Comparison of the ABCD Rule of Dermatoscopy and a New 7-Point Checklist. *Arch. Dermatol.* **1998**, *134*, 1563–1570. [CrossRef]
44. Ha, Q.; Liu, B.; Liu, F. Identifying Melanoma Images using EfficientNet Ensemble: Winning Solution to the SIIM-ISIC Melanoma Classification Challenge. *arXiv* **2020**, arXiv:2010.05351. [CrossRef]
45. Cohen, J. A coefficient of agreement for nominal scales. *Educ. Psychol. Meas.* **1960**, *20*, 37–46. [CrossRef]
46. Landis, J.R.; Koch, G.G. The measurement of observer agreement for categorical data. *Biometrics* **1977**, *33*, 159–174. [CrossRef]
47. Lancaster, K.; Zouridakis, G. Asymmetry Measures of Dermoscopic Images for Automated Melanoma Detection. In *Proceedings of the 2023 IEEE 7th Portuguese Meeting on Bioengineering (ENBENG)*; IEEE: New York, NY, USA, 2023; pp. 151–154.
48. Rotemberg, V.; Kurtansky, N.; Betz-Stablein, B.; Caffery, L.; Chousakos, E.; Codella, N.; Combalia, M.; Dusza, S.; Guitera, P.; Gutman, D.; et al. A patient-centric dataset of images and metadata for identifying melanomas using clinical context. *Sci. Data* **2021**, *8*, 34. [CrossRef]
49. Queen, C.M.; Hu, R.; Zouridakis, G. Towards the development of reliable and economical mHealth solutions: A methodology for accurate detection of Buruli ulcer for hard-to-reach communities. *Front. Trop. Dis.* **2023**, *3*, 1031352. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.