Health and Technology
https://doi.org/10.1007/s12553-025-00971-x

ORIGINAL PAPER

Advancing synthetic data for dermatology: GAN comparison with multi-metric and expert validation approach

Alessio Luschi[1] · Linda Tognetti[2] · Alessandra Cartocci[2] · Gabriele Cevenini[1] · Pietro Rubegni[2] · Ernesto Iadanza[1,3]

Received: 14 March 2025 / Accepted: 16 April 2025

Abstract

Background Malignant melanoma (MM) is the most aggressive skin cancer, requiring early diagnosis for better outcomes. While deep learning models have shown promise in dermatological image analysis, their performance is constrained by limited training data. Generative Adversarial Networks (GANs) offer a solution by generating synthetic images for data augmentation. However, assessing their clinical reliability remains difficult as automated metrics may not fully capture visual realism or clinical usability.

The Objective This study presents a comprehensive framework for creating high-quality synthetic dermatoscopic pictures of MM lesions, as well as a holistic validation methodology that uses quantitative metrics and qualitative dermatologist assessment to provide a full clinical evaluation of the generated images. Three GAN architectures (DCGAN, StyleGAN2, and StyleGAN3-t) are explored. Lesions on the face, palms, and soles are excluded due to their unique dermoscopic patterns.

Materials and Methods A dataset of 1,774 dermatoscopic body MM images was used to train the models, assessed with Fréchet Inception Distance (FID), Kernel Inception Distance (KID), precision, and recall. Afterwards, a panel of 17 dermatologists with different levels of expertise assessed image quality using a 7-point Likert scale, with accuracy, sensitivity, specificity, and inter-rater agreement analysed.

Results StyleGAN2 achieved the best image fidelity (FID: 18.89, KID: 0.0025), while StyleGAN3-t demonstrated stable but slower convergence. Both StyleGAN models outperformed DCGAN in diversity and fidelity. The validation study showed that StyleGAN2-generated images were often indistinguishable from real ones, reflected in low specificity and accuracy values among evaluators.

Conclusions The study highlights the effectiveness of GANs in generating high-quality synthetic images, proposing a validation framework that integrates expert assessments with state-of-the-art quantitative metrics. This approach advances standardisation in GAN evaluation, ensuring synthetic images are clinically relevant for dermatological AI applications.

Keywords Generative AI · Dermatology · GAN

1 Introduction

Malignant melanoma (MM) is the most aggressive form of skin cancer, posing a significant burden on public health [1]. According to the International Agency for Research on Cancer, more than 330,000 new cases were reported worldwide in 2022, resulting in approximately 58,000 deaths [2]. In the United States alone, people experience an average of 1.6 skin diseases per year, highlighting the widespread nature of these conditions [3, 4]. In particular, dermatological concerns account for almost 20% of all general practitioner (GP) visits, making them one of the most common reasons to seek primary care [5].

Dermoscopy considerably improved early melanoma detection; successful application, however, necessitates wide-ranging training and is highly dependent on the user's expertise [6]. Histopathology remains the gold standard for the diagnosis of skin cancer: it is based on the identification of several specific biomarkers along with a degree of cellular atypia. The accuracy of this method is high, but it requires significant expertise and is very time-consuming [7]. An effective treatment for MM did not exist before 2011 [8].

Extended author information available on the last page of the article

Published online: 25 April 2025

Furthermore, there is no conclusive evidence that melanoma screening reduces morbidity or mortality [9]. Taking into account these factors, it was anticipated that a large increase in melanoma-related deaths would follow the increase in MM cases before 2011. However, mortality rates have remained stable. This trend has persisted since the year 2011. This suggests that a significant portion of diagnoses could be due to overdiagnosis or misdiagnosis [10]. This generates many diagnostic issues and increases the number of unnecessary surgeries or incorrect treatments, with a subsequent impact on aesthetic outcomes and healthcare costs.

Artificial Intelligence (AI) has been widely regarded as a promising tool to address these challenges. In particular, Deep Learning (DL) has the potential to enhance diagnostic accuracy while reducing both misdiagnoses and overdiagnoses [11]. Convolutional Neural Networks (CNNs), known for their ability to process grid-like data structures, have become the gold standard in medical image analysis [12, 13]. Dermatology, in particular, is well-suited for AI applications as it heavily relies on dermatoscopic imaging, patient history data, and descriptive pathology reports. However, CNNs require large volumes of training data, which presents a major limitation in achieving reliable performance in clinical practice, primarily due to the unbalanced distributions of pathologies, the limited representativeness of rare clinical conditions, and the insufficient variability in factors such as ethnicity and image quality.

A promising solution to this data scarcity issue is the use of synthetic data, which can enhance domain knowledge through synthetic augmentation [14]. Studies have shown that synthetic datasets can achieve even better visual performance than real datasets [15]. Generative Adversarial Networks (GANs) are particularly effective at generating high-quality synthetic images for data augmentation [16]. GANs operate through a zero-sum game between two competing neural networks, where one network's success equates to the other's failure. Given a training dataset, this method learns to produce new data with statistical properties similar to the original set.

However, the use of synthetic data in healthcare applications also raises concerns such as the risk of data leakage, dependency on imputation models, bias amplification, low interpretability, and an absence of robust methods for auditing data quality [17, 18]. Even if GANs are proven to reduce re-identification risks, thus holding significant potential for privacy-preserving data sharing in healthcare, challenges persist in preserving temporal correlations and aligning with regulatory frameworks [19, 20].

1.1 Related works

The paper by Goodfellow et al. [16] sets a very important milestone in the modern machine learning landscape, opening new avenues for unsupervised learning. Deep Convolutional GAN (DCGAN) continued building on this idea by applying the principles of a convolutional network to produce 2D images successfully [21]. DCGANs integrate the CNN architectures into GANs, making them particularly effective for image generation tasks. In DCGANs, the Generator employs transposed convolutions to generate high-level data distributions, while the Discriminator uses convolutional layers to classify the data.

Cycle-Consistent Generative Adversarial Network (CycleGAN) is a GAN framework designed to transfer features between images [22]. CycleGAN is best suited for unpaired image-to-image translation tasks when the input and output images are not connected. CycleGAN is used for a variety of image-to-image translation challenges, such as style transfer, domain adaptation, and image enhancement.

Another GAN architecture is the StyleGAN developed by NVIDIA [23]. StyleGAN is designed primarily to make high-quality, photo-realistic images. Unlike traditional GANs, StyleGAN provided various unique ways to improve image production and greater control over specific attributes. Unlike conventional GAN, StyleGAN uses a progressive growing technique that allows the generator and discriminator networks to grow in size and sophistication throughout training. StyleGAN may produce images of higher quality (up to 1024x1024 pixels) because of progressive growth.

Other common GAN architectures used for image synthesis are Conditional GAN (CGAN) [24], Pix2Pix [25], Wasserstein GAN (WGAN) [26], and DiscoGAN [27]. CGAN allows for control over the generated output by conditioning the generation process on additional information; the Pix2Pix model is based on a CGAN where a target image is generated based on a given input image; WGAN introduces a loss function based on the Wasserstein distance, which provides smoother and more meaningful gradients preventing unstable convergence; finally, DiscoGAN generates images of products in one domain given an image in another domain, thus having the ability to learn cross-domain relations given unsupervised data.

However, each GAN architecture is proven to suffer from inadequate learning caused by two common aspects known as convergence failure (a scenario when GAN training fails to converge without even producing any meaningful content) and mode collapse (the generator focuses on producing a limited set of data patterns that deceive the discriminator). Some common solutions that can help mitigate these failures are the Adaptive Moment Estimation (Adam) optimisation algorithm for gradient descent [28], Spectral Normalization as opposed to Batch Normalization as a weight regularisation technique [29], use of learning rate schedulers [30], and Label Smoothing [31].

The evaluation of GANs has emerged as a hot topic in AI, particularly in terms of generated image realism

and quality. Several state-of-the-art quantitative metrics are widely utilised to evaluate the quality of images produced by GANs, such as the Inception Score (IS) [32], the Fréchet Inception Distance (FID) [33], the Kernel Inception Distance (KID) [34], precision and recall [35]. α-precision, β-recall and authenticity [36]. However, all existing evaluation approaches face several obstacles, the most significant of which is the lack of a common standard capable of accurately capturing the intricacies of distinct generative models [19]. In such a scenario, the subjective nature of human judgment can play a crucial role in image quality assessment. While automated metrics provide useful information, human evaluations frequently capture aspects of realism and quality that quantitative tools may miss. Studies have shown a substantial correlation between human qualitative assessments and traditional metrics, indicating that including human input can significantly improve evaluation accuracy [19, 37].

A few studies have already explored the application of GANs in dermatology. However, they generally obtained poor results in terms of image quality and resolution. Moreover, no validation standard framework is presented as they solely rely on quantitative metrics (e.g., FID and IS) that present the previously highlighted limitations.

The study by Furger et al. explored the application of GANs to three different types of images showing full hands, skin lesions, and varying degrees of eczema, obtaining unrealistic and unusable results for skin lesions such as MM, with a FID score of 599.6 for skin eczema [38].

The work by Borghesi et al. proposes the generation of synthetic skin images of a resolution of 256x256 pixels with a diffusion model and with a scarce training set [39].

Abbasi et al. aim to demonstrate the potential of the pre-trained VGG-16 deep learning model [40] to accurately differentiate between real and GAN-generated medical images. The results show that the model achieves an accuracy of 99.82% in discriminating between real and synthetic images. However, no information is given about the performance of the proposed GAN, which prevents any assessment of the quality of the synthetic images to be discussed [41].

1.2 The objective

This manuscript presents a complete framework for generating high-quality synthetic dermatoscopic images of MM lesions, and a holistic validation protocol that leverages quantitative metrics and qualitative assessment to offer a full clinical evaluation of the generated images. The proposed work will help augment current datasets, offering a valid tool to improve the performance of deep learning classification models indirectly. Unlike recently proposed approaches [38, 39, 41], this study exclusively focuses on MM lesions on the body. Lesions from the face, palms, and soles are deliberately excluded, as these areas exhibit distinct dermoscopic patterns. The anatomical location of MM lesions is proven to influence the dermoscopic appearance of MM lesions significantly [42]. Therefore, training a GAN on images from different regions may result in unrealistic and clinically unhelpful synthetic images. A validation protocol is also presented to provide a complete and effective clinical evaluation and a more nuanced understanding of the quality and diversity of the generated images. Three GAN models (a custom-developed DCGAN, a fine-tuned StyleGAN2, and a fine-tuned StyleGAN3-t) were trained on a curated set of body MM images sourced from multiple databases. The proposed assessment integrates state-of-the-art quantitative metrics with qualitative assessments performed by a panel of dermatologists with different levels of expertise.

2 Materials and methods

1,430 dermatoscopic images were collected from two open-source databases: 970 from the HAM10000 (Human Against Machine) [43] and 460 from the ISIC 2020 Challenge Dataset [44]. These datasets were complemented with 344 additional dermatoscopic scans from the multicentric IdScore_Body dataset [45]. The combination of the three databases yielded a total of 1,774 dermatoscopic scans exclusively linked to body MM lesions.

The first proposed GAN model is based on the DCGAN architecture for generating 256x256 pixels synthetic images of skin alterations. It implements the following guidelines as introduced by Radford et al. [21]:

- Replacing any pooling layers with stride convolutions (Discriminator) and fractional-stride convolutions (Generator).
- Using batch normalization in both the Generator and the Discriminator.
- Removing fully connected hidden layers for deeper architectures.
- Using Rectified Linear Unit (ReLU) activation function in the Generator for all layers except for the output, for which *Tanh* is used.
- Using *LeakyReLU* activation function in the Discriminator for all layers, except for the output, which uses *Sigmoid*.

The designed architecture of the Discriminator also exploits spectral normalization, learning rate schedulers, and label smoothing to enhance performance and stabilize the training process. The last technique refers to changing the usual value of the classification labels, typically 1 or 0, to another value close to the original. By reducing the value of the assigned label, the Discriminator gets "less confident"

since the assigned probabilities at the output layer are lower in comparison with what happens by using the original label value. In this implementation, the real images, originally labelled as 1, have been labelled with a value of 0.9. The developed DCGAN model is publicly available on a GitHub repository [46].

The second and third tested GAN models exploit NVIDIA's StyleGAN2 and StyleGAN3-t (translation equivalent) architectures. In the latter, the previous generator architecture has been modified by the authors to get greater control over the image synthesis process by injecting noise with a constant learning input and by adjusting the image style in each convolutional layer based on the latent code that enhances the image attributes [47]. Both models employ adaptive discriminator augmentation (ADA) to generate synthetic images with a resolution of 512x512 pixels. ADA is a technique that dynamically adjusts the augmentation applied to training data based on the model's needs. Traditional data augmentation (e.g., rotation, scaling, translation, colour adjustments) can lead to undesirable artefacts and noisy results. ADA monitors the training process and adapts the intensity of the augmentation to balance performance and prevent problems such as mode collapse or overfitting by selectively applying augmentations with a given probability. This ensures that the generator learns to produce natural images while still benefiting from the diversity introduced by augmentation [47]. The probability is regulated by the hyperparameter p, in the range of 0 to 1. StyleGANs were fine-tuned starting from the pre-trained *stylegan2-ffhq-512x512*. When the training set is minimal (as in our situation), this fine-tuning allows for quick and high-quality results. Finally, because the dataset is horizontally symmetrical, data augmentation using random x-flips was performed, doubling the number of images. The DCGAN was trained for 500 epochs while the StyleGANs for 15,000 training iterations, or until diverging instead of converging to a stable equilibrium on the selected images. The hyperparameter selection for the DCGAN has been optimized with a trial-and-error approach, testing different combinations and assessing the final performances. Hyperparameters for the StyleGAN2 and StyleGAN3-t architectures have been set according to previous studies [23, 47].

Table 1 shows the value of the hyperparameters chosen for training the three GAN architectures.

The performance of each model was assessed using quantitative state-of-the-art metrics to assess fidelity (the realism of the synthetic data) and diversity (the coverage of the real data distribution): FID, KID, precision, and recall.

FID measures the difference between the real and generated image distributions by comparing their feature representations extracted from the *InceptionV3* deep neural network. It assumes the features follow a multivariate Gaussian distribution and it is calculated as per Eq. 1.

$$FID = \|\mu_r - \mu_g\|^2 + \mathrm{Tr}(\Sigma_r + \Sigma_g - 2\sqrt{\Sigma_r \Sigma_g}) \quad (1)$$

- μ_r, Σ_r are the mean and covariance of the real image features,
- μ_g, Σ_g are the mean and covariance of the generated image features,
- Tr denotes the trace of a matrix.

Usually, a FID score lower than 15 denotes excellent and high-quality images with a variation due to the dataset complexity. A score greater than 100 is instead associated with very low-quality images, possibly due to mode collapse or artefacts [33].

KID also measures how closely the generated images resemble real images by comparing their feature representations extracted from the *InceptionV3* model. It differs from FID in how it computes this similarity. KID is based on the

Table 1 Initial hyperparameters for the tested GAN models

Parameter	DCGAN	StyleGAN2	StyleGAN3-t
Output resolution (pixels)	256x256	512x512	512x512
Batch size	16	16	16
Training epochs/iterations	500	15,000	15,000
Latent vector size	128	512	512
Learning rate (Generator)	0.0002	0.0025	0.0025
Learning rate (Discriminator)	0.00005	0.0025	0.002
$\beta 1$ (Adam optimizer)	0.5	0.0	0.0
$\beta 2$ (Adam optimizer)	0.999	0.99	0.99
Decay epoch (LambdaLR scheduler)	100	–	–
p	–	0.7	0.7
Transfer learning	N.A.	✓	✓
Random x-flips	N.A.	✓	✓

Maximum Mean Discrepancy (MMD), a statistical test that measures the difference between two probability distributions using a polynomial kernel function (Eq. 2).

$$KID = \mathrm{MMD}^2(\mathbb{P}_r, \mathbb{P}_g) \tag{2}$$

$\mathbb{P}_r$ and $\mathbb{P}_g$ represent the feature distributions of real and generated images, respectively. Low KIDs mean the generated images are similar to real ones. A KID score less than 0.01 is associated with high quality, while a value greater than 0.1 highlights poor-quality generations, likely mode collapse or artefacts [34].

The implemented precision and recall are defined in the work by [35] and they are used to assess the fidelity and diversity of the generated images, respectively. Real and generated samples are drawn and embedded into a high-dimensional feature space using a pre-trained d VGG-16 [40] classifier network. Vectors of the real and generated images are denoted by ϕ_r and ϕ_g, while the corresponding sets of feature vectors are by Φ_r and Φ_g. Precision is quantified by querying for each generated image whether the image is within the estimated manifold of real images (Eq. 3). Symmetrically, recall is calculated by querying for each real image whether the image is within the estimated manifold of the generated image (Eq. 4)

$$\mathrm{precision}(\Phi_r, \Phi_g) = \frac{1}{|\Phi_g|} \sum_{\phi_g \in \Phi_g} f(\phi_g, \Phi_r) \tag{3}$$

$$\mathrm{recall}(\Phi_r, \Phi_g) = \frac{1}{|\Phi_r|} \sum_{\phi_r \in \Phi_r} f(\phi_r, \Phi_g) \tag{4}$$

A precision score close to 1 denotes a high fidelity of the generated images to the real ones, while a recall score close to 1 proves a high diversity of the synthetic image sample.

2.1 Validation protocol

A user study on selected physicians was also designed to evaluate the perceived quality and diversity of the generated samples and ensure a strong validation of the best-performing model, chosen according to the lowest FID score. Specifically, 17 dermatologists distributed within three different levels of expertise (five with <4 years, six with 4-8 years, and six with >8 years) [48] were enrolled from three different Italian centres (the University of Siena, the University Luigi Vanvitelli in Caserta, and the University Sapienza in Rome). The validation dataset was made with 25 generated synthetic images and the same number of randomly selected real images among the training dataset. Images had the same resolution and were shuffled and anonymised to build a randomised balanced dataset of real and synthetic images with a total of 50 images. Dermatologists were unaware of each image's origin to ensure unbiased assessment. A custom web application deployed on a publicly accessible server [49] was used to administer the questionnaires (expertise and realism assessment) to the selected dermatologists and store the answers. An initial assessment was requested from the dermatologists to discriminate real/synthetic images. Afterward, each image was also assessed on three attributes (image quality, visual and colour realism, and confidence level of the evaluation) with a 7-point Likert scale. Descriptive statistics were used to describe the results of the qualitative validation.

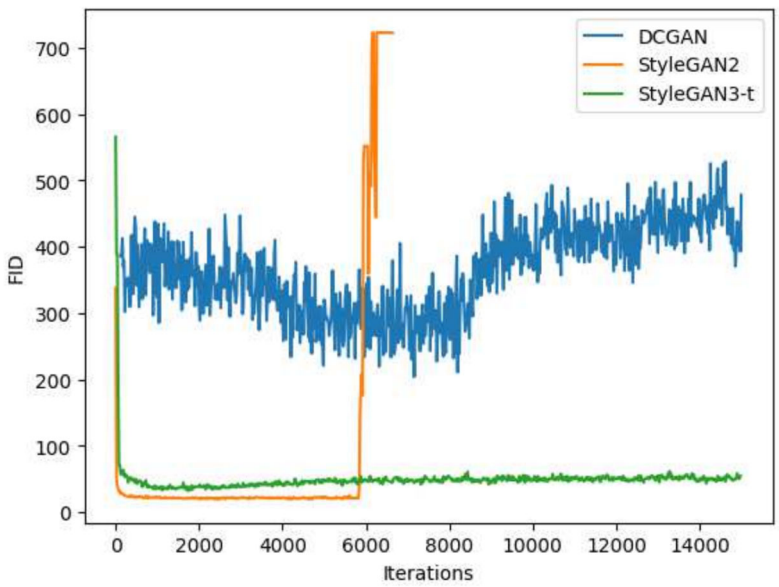


Fig. 1 Evolution of FID values of the three GAN models during training

Table 2 Value of the employed metrics (FID, KID, Precision, and Recall) for the three different GAN models evaluated individually at the iteration with the lowest FID

Model	FID	KID	Precision	Recall
DCGAN	203.77	0.0660	**0.7500**	**0.8125**
StyleGAN2	**18.89**	**0.0025**	0.7408	0.2604
StyleGAN3-t	31.73	0.0220	0.6616	0.3619

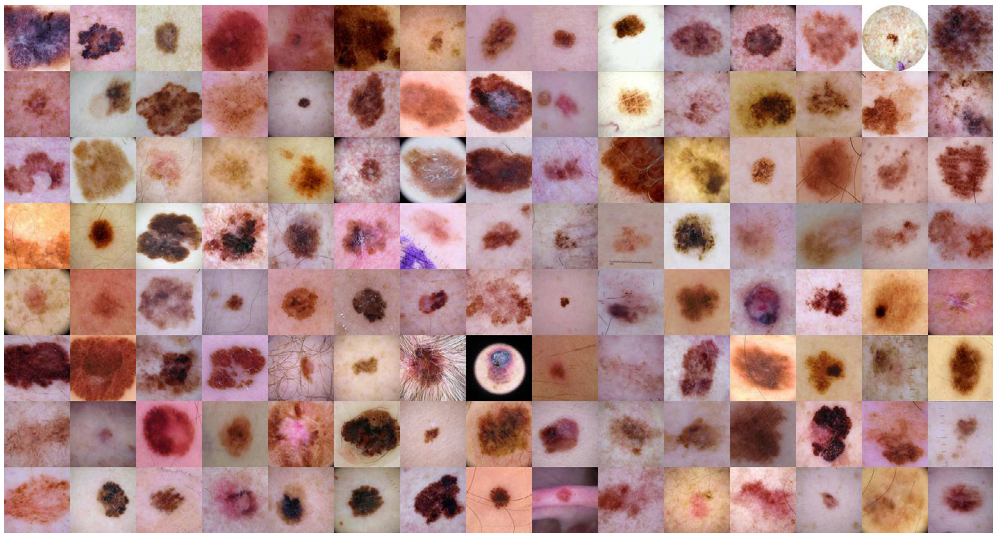
The single best metrics are highlighted in bold

Specifically, accuracy, sensitivity, and specificity were evaluated on the whole assessment and stratified on the evaluators' expertise and confidence level, image quality, and visual and colour realism. Fleiss' Kappa was used to measure the concordance of the evaluation on the same image, while Krippendorf's Alpha was calculated to assess the inter-rater agreement of the evaluation on the parameters scored with the 7-point Likert scale (image quality, visual and colour realism, and confidence level).

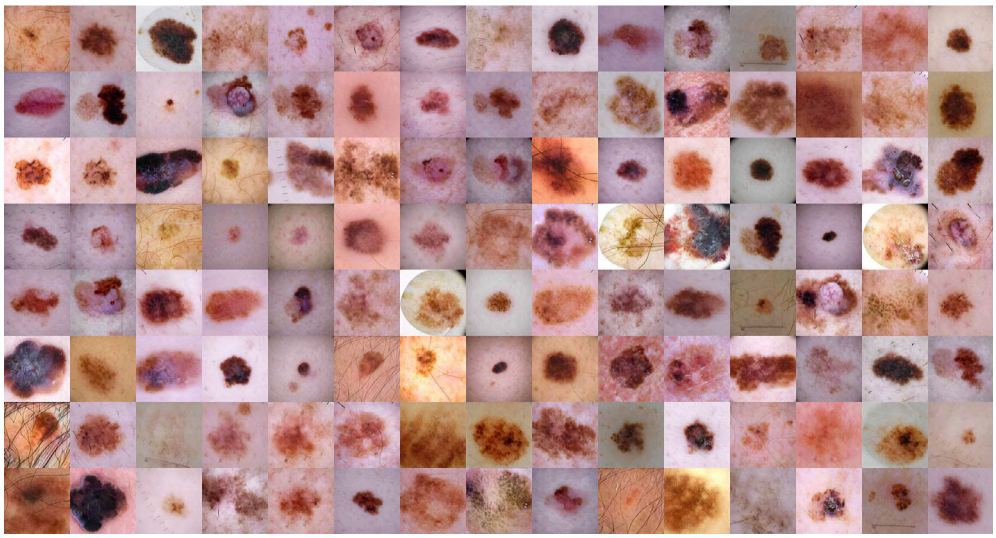
3 Results

3.1 Quantitative assessment

Figure 1 shows the trend of the FID during the training of the three GAN models.



(a) Sample of 120 real dermatoscopic images.



(b) Sample of 120 synthetic dermatoscopic images.

Fig. 2 Comparison between real (a) and synthetic (b) dermatoscopic images of MM body lesions. Synthetic images are generated by the snapshot of the StyleGAN2 model trained for 4,620 iterations which scored the lowest FID of 18.89

The DCGAN reaches the lowest FID value of 203.77 after 7,160 iterations. StyleGAN2 scores the smallest FID of 18.89 after 4,620 iterations, while StyleGAN3-t has a minimum FID of 31.73 after 1,740 iterations (Table 2).

The snapshot of the StyleGAN2 model, taken after 4,620 iterations, was then selected as the best-performing model. Figure 2b shows a sample of 120 synthetic images generated by the best-performing model compared to 120 real ones (Fig. 2a) randomly selected from the initial set of 1,774.

3.2 Qualitative assessment

The results of the statistical analysis on the qualitative validation are reported in Table 3. A true-positive response is identified by a real image identified as real by the participants; a false-positive whenever a synthetic image is identified as real; a true-negative whenever a synthetic image is labelled as synthetic; a false-negative each time a real image is classified as synthetic.

Results from the parameters assessed with the 7-point Likert scale (image quality, visual and colour realism, and confidence level) were grouped into three subsets: low (1-3), medium (4-5), and high (6-7).

Statistical analysis reveals a consistently low level of accuracy across different stratifications, with a slightly higher value (0.74; IC: 95% = 0.65-0.82) observed among those who reported a high confidence level in their assessments. The sensitivity for the entire assessment is 0.73; IC: 95% = 0.69-0.78, with higher values recorded for images of medium to high quality, high visual realism, and high confidence of the evaluator. The lowest sensitivity (0.11; IC: 95% = 0.04-0.24) is associated with images perceived as having low visual realism. The specificity has a consistent trend with low values across the various stratifications (with the lowest value of 0.18; IC: 95% = 0.09-0.30 for the evaluators that assessed a High Visual Realism of the image), except for the participants that assessed a low visual realism of the image (0.97; IC: 95% = 0.90-0.99). Intra-rater concordance measured with the Fleiss' Kappa and the Krippendorf's Alpha is generally very low except for evaluators who expressed a high level of confidence in their assessment (0.53).

4 Discussion

Figure 1 highlights that the DCGAN model exhibits a significantly higher FID score and a more unstable trend compared to the StyleGAN models. In particular, the StyleGAN2 model shows a rapid decline in FID, followed by divergence after approximately 6,000 iterations, whereas the StyleGAN3-t model exhibits a more stable but slower convergence. This difference can be attributed to both architectural variations among the GAN models and the use of ADA and transfer-learning in the StyleGAN models, which allows a much faster convergence toward high-quality image generation while preventing overfitting. Furthermore, the KID scores of both StyleGAN3-t and StyleGAN2 are substantially lower than that of DCGAN, reinforcing their superior diversity and fidelity. However, precision values remain comparable across all models, and recall is even higher for DCGAN, suggesting, paradoxically, a greater diversity for this model. These findings underscore the necessity of employing multiple evaluation metrics when assessing synthetic image quality across different domains, as relying solely on precision and recall could have led to misleading conclusions.

The low accuracy of the qualitative assessment indicates an overall high quality of the generated synthetic images,

Table 3 Statistical analysis on the results of the qualitative validation assessment

	Accuracy	Sensitivity	Specificity	Fleiss' Kappa	Krippendorf's Alpha
Whole Assessment	0.64 [0.61-0.68]	0.73 [0.69-0.78]	0.56 [0.51-0.60]	0.183	–
Low expertise [<4 years]	0.59 [0.53-0.65]	0.67 [0.59-0.74]	0.51 [0.43-0.60]	0.074	–
Medium expertise [4-8 years]	0.64 [0.59-0.70]	0.78 [0.71-0.84]	0.51 [0.42-0.59]	0.254	–
High expertise [>8 years]	0.71 [0.65-0.77]	0.76 [0.68-0.83]	0.66 [0.57-0.75]	0.225	–
Low Image Quality	0.64 [0.58-0.69]	0.48 [0.39-0.58]	0.74 [0.67-0.80]	–	0.047
Medium Image Quality	0.65 [0.60-0.69]	0.82 [0.77-0.87]	0.44 [0.37-0.51]	–	0.118
High Image Quality	0.64 [0.55-0.73]	0.85 [0.74-0.93]	0.35 [0.22-0.51]	–	−0.040
Low Visual Realism	0.67 [0.58-0.75]	0.11 [0.04-0.24]	0.97 [0.90-0.99]	–	−0.025
Medium Visual Realism	0.63 [0.59-0.67]	0.76 [0.70-0.81]	0.50 [0.44-0.56]	–	0.121
High Visual Realism	0.68 [0.60-0.75]	0.97 [0.91-0.99]	0.18 [0.09-0.30]	–	0.038
Low Confidence Level	0.60 [0.49-0.70]	0.50 [0.32-0.68]	0.66 [0.52-0.78]	–	0.032
Medium Confidence Level	0.63 [0.60-0.67]	0.72 [0.67-0.77]	0.55 [0.50-0.60]	–	0.155
High Confidence Level	0.74 [0.65-0.82]	0.89 [0.80-0.95]	0.43 [0.27-0.61]	–	0.53

Upper and lower values of the 95% confidence interval are reported in brackets for accuracy, sensitivity, and specificity

as they successfully "deceive" experts, leading to frequent misclassification. The sensitivity scores suggest that evaluators generally have strong expertise as they can accurately identify real dermatoscopic images within the dataset, especially when these images exhibit high quality and realism. However, the consistently low specificity values highlight the challenge physicians face in distinguishing synthetic images from real ones. Dermatologists were always able to classify synthetic images with low visual realism correctly. The overall low concordance among evaluators (with the only exception of evaluators expressing a high level of confidence in their assessments) further supports the high quality of the synthetic images.

5 Conclusions

The study evaluated the effectiveness of GANs in generating high-quality synthetic dermatoscopic images of body MM lesions, comparing DCGAN, StyleGAN2, and StyleGAN3-t. The findings demonstrate that StyleGAN models, particularly StyleGAN2, outperform DCGAN in producing realistic and diverse images. The quantitative evaluation showed that StyleGAN2 achieved the lowest FID (18.89) and KID (0.0025) scores, indicating superior image fidelity and diversity, while StyleGAN3-t exhibited more stable but slower convergence. The identified discrepancy among precision/recall and FID/KID scores of the evaluated models, underscores the necessity of using multiple evaluation metrics to obtain a comprehensive assessment of synthetic image quality. To address the limitations of relying only on automated metrics, this study introduced a human validation protocol involving expert dermatologists. The results demonstrated that evaluators struggled to identify synthetic images from real ones regardless of expertise, as evidenced by low specificity and accuracy values. The proposed qualitative validation protocol represents a significant advancement in the assessment of GAN-generated images. This approach provides a more accurate and holistic evaluation of GAN performance by integrating expert evaluations with established quantitative metrics. The proposed methodology enhances the reliability of synthetic image validation and paves the way for a standardised assessment framework for GAN-generated medical images.

Acknowledgements The authors would like to thank the panel of dermatologists who assessed the generated synthetic images at the University of Siena, University Luigi Vanvitelli in Caserta, and University Sapienza in Rome.

Funding Open access funding provided by Università degli Studi di Siena within the CRUI-CARE Agreement. The authors declare that no funds, grants, or other support were received during the preparation of this manuscript.

Declarations

Ethical Approval The dataset utilized human dermatological scans that were procured via the open-access datasets HAM10000 and ISIC, which provides de-identified samples. The IDScore_Body project were approved by the Ethical Committee "Comitato Etico Regionale per la Sperimentazione Clinica della Regione Toscana" - Approval Code: 16801 - Approval Date: 03/16/2020.

Consent to Participate N.A.

Consent to Publish N.A.

Competing interests The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Arnold M, Singh D, Laversanne M, Vignat J, Vaccarella S, Meheus F, Cust AE, Vries E, Whiteman DC, Bray F. Global burden of cutaneous melanoma in 2020 and projections to 2040. JAMA Dermatol. 2022;158(5):495–503.
2. Lowe GC, Saavedra A, Reed KB, Velazquez AI, Dronca RS, Markovic SN, Lohse CM, Brewer JD. Increasing incidence of melanoma among middle-aged adults: An epidemiologic study in olmsted county, minnesota. Mayo Clin Proc. 2014;89(1):52–9.
3. Lim HW, Collins SAB, Resneck JS, Bolognia JL, Hodge JA, Rohrer TA, Van Beek MJ, Margolis DJ, Sober AJ, Weinstock MA, Nerenz DR, Smith Begolka W, Moyano JV. The burden of skin disease in the united states. J Amer Acad Dermatol. 2017;76(5):958–9722.
4. Schofield JK, Fleming D, Grindlay D, Williams H. Skin conditions are the commonest new reason people present to general practitioners in England and Wales. Br J Dermatol. 2011;165(5):1044–50.
5. Lowell B, Froelich C, Federman D, Kirsner R. Dermatology in primary care: Prevalence and patient disposition. J Am Acad Dermatol. 2001;45:250–5.
6. Williams NM, Rojas KD, Reynolds JM, Kwon D, Shum-Tien J, Jaimes N. Assessment of Diagnostic Accuracy of Dermoscopic Structures and Patterns Used in Melanoma Detection: A Systematic Review and Meta-analysis. JAMA Dermatol. 2021;157(9):1078–88.
7. Fischman S, Pérez Anker J, Tognetti L, Naro A, Suppa M, Cinotti E, Viel T, Monnier J, Rubegni P, Del Marmol V, Malvehy J, Puig S, Dubois A, Perrot J. Non-invasive scoring of cellular atypia in keratinocyte cancers in 3d lc-oct images using deep learning. Sci Rep. 2022;12:481.
8. Kahlon N, Doddi S, Yousif R, Najib S, Sheikh T, Abuhelwa Z, Burmeister C, Hamouda DM. Melanoma treatments and mor-

tality rate trends in the us, 1975 to 2019. JAMA Netw Open. 2022;5(12):2245269–2245269.
9. Johansson M, Brodersen J, Gøtzsche P, Jørgensen K. Screening for reducing morbidity and mortality in malignant melanoma. Cochr Datab System Rev. 2019;(6).
10. Bjørch MF, Gram EG, Brodersen JB. Overdiagnosis in malignant melanoma: a scoping review. BMJ Evidence-Based Medicine. 2024;29(1):17–28.
11. Li Z, Koban KC, Schenck TL, Giunta RE, Li Q, Sun Y. Artificial intelligence in dermatology image analysis: Current developments and future trends. J Clin Med. 2022;11(22):6826.
12. Cartocci A, Luschi A, Tognetti L, Cinotti E, Farnetani F, Lallas A, Paoli J, Longo C, Moscarella E, Tiodorovic D, Stanganelli I, Suppa M, Dika E, Zalaudek I, Pizzichetta MA, Perrot JL, Cevenini G, Iadanza E, Rubegni G, Kittler H, Tschandl P, Rubegni P. Comparative analysis of ai models for atypical pigmented facial lesion diagnosis. Bioengineering. 2024;11(10):1036.
13. Tognetti L, Miracapillo C, Leonardelli S, Luschi A, Iadanza E, Cevenini G, Rubegni P, Cartocci A. Deep learning techniques for the dermoscopic differential diagnosis of benign/malignant melanocytic skin lesions: From the past to the present. Bioengineering. 2024;11(8):758.
14. Jain S, Seth G, Paruthi A, Soni U, Kumar G. Synthetic data augmentation for surface defect detection and classification using deep learning. J Intell Manuf. 2022;33(4):1007–20.
15. Wang X, Xie L, Dong C, Shan Y. Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In: 2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW); 2021. p. 1905–1914.
16. Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y. Generative adversarial networks. Commun ACM. 2020;63(11):139–44.
17. Gonzales A, Guruswamy G, Smith S. Synthetic data in health care: A narrative review. PLOS Digital Health. 2023;2:0000082.
18. Giuffrè M, Shung D. Harnessing the power of synthetic data in healthcare: innovation, application, and privacy. npj Digit Med. 2023;6:186.
19. Ghiurău D, Popescu DE. Distinguishing Reality from AI: Approaches for Detecting Synthetic Content. Computers. 2025;14(1):1.
20. Liu Y, Acharya UR, Tan JH. Preserving privacy in healthcare: A systematic review of deep learning approaches for synthetic data generation. Comput Methods Programs Biomed. 2025;260:108571.
21. Radford A, Metz L, Chintala S. Unsupervised representation learning with deep convolutional generative adversarial networks. ArXiv 2015. https://doi.org/10.48550/arXiv.1511.06434
22. Zhu J-Y, Park T, Isola P, Efros AA. Unpaired image-to-image translation using cycle-consistent adversarial networks. In: 2017 IEEE International Conference on Computer Vision (ICCV); 2017. p. 2242–2251.
23. Karras T, Laine S, Aila T. A Style-Based Generator Architecture for Generative Adversarial Networks. IEEE Trans Patt Anal Mach Intell. 2021;43(12):4217–28.
24. Mirza M, Osindero S. Conditional generative adversarial nets. ArXiv 2014. https://doi.org/10.48550/arXiv.1411.1784
25. Isola P, Zhu J-Y, Zhou T, Efros AA. Image-to-image translation with conditional adversarial networks. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2017. p. 5967–5976.
26. Arjovsky M, Chintala S, Bottou L. Wasserstein generative adversarial networks. In: Proceedings of the 34th International Conference on Machine Learning - Volume 70. ICML'17. JMLR.org, Sydney, NSW, Australia; 2017. p. 214–223.
27. Kim T, Cha M, Kim H, Lee JK, Kim J. Learning to discover cross-domain relations with generative adversarial networks. In: Proceedings of the 34th International Conference on Machine Learning - Volume 70. ICML'17. JMLR.org, Sydney, NSW, Australia; 2017. p. 1857–1865.
28. Kingma DP, Ba J. Adam: A method for stochastic optimization. ArXiv 2015. https://doi.org/10.48550/arXiv.1412.6980
29. Miyato T, Kataoka T, Koyama M, Yoshida Y. Spectral normalization for generative adversarial networks. ArXiv 2018. https://doi.org/10.48550/arXiv.1802.05957
30. Li K, Kang D-K. Enhanced generative adversarial networks with restart learning rate in discriminator. Appl Sci. 2022;12(3):1191.
31. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. Rethinking the inception architecture for computer vision. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2016. p. 2818–2826.
32. Salimans T, Goodfellow I, Zaremba W, Cheung V, Radford A, Chen X. Improved techniques for training gans. In: Proceedings of the 30th International Conference on Neural Information Processing Systems. NIPS'16. Curran Associates Inc., Red Hook, NY, USA; 2016. p. 2234–2242.
33. 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 31st International Conference on Neural Information Processing Systems. NIPS'17. Curran Associates Inc., Red Hook, NY, USA; 2017. p. 6629–6640.
34. Bińkowski M, Sutherland D, Arbel M, Gretton A. Demystifying mmd gans. ArXiv 2018. https://doi.org/10.48550/arXiv.1801.01401
35. Kynkäänniemi T, Karras T, Laine S, Lehtinen J, Aila T. Improved precision and recall metric for assessing generative models. In: Proceedings of the 33rd International Conference on Neural Information Processing Systems. Curran Associates Inc., Red Hook, NY, USA; 2019. p. 353–362.
36. Alaa A, Breugel B, Saveliev E, Schaar M. How faithful is your synthetic data? sample-level metrics for evaluating and auditing generative models. ArXiv 2021. https://doi.org/10.48550/arXiv.2102.08921
37. Li R, Jin X, Ji H. Real: Realism evaluation of text-to-image generation models for effective data augmentation. ArXiv 2025. https://doi.org/10.48550/arXiv.2502.10663
38. Furger F, Amruthalingam L, Navarini A, Pouly M. Applications of generative adversarial networks to dermatologic imaging. In: Schilling F-P, Stadelmann T, editors. Artificial Neural Networks in Pattern Recognition. Cham: Springer; 2020. p. 187–99.
39. Borghesi A, Calegari R. In: Shaban-Nejad A, Michalowski M, Bianco S, editors. Generation of Clinical Skin Images with Pathology with Scarce Data. Springer, Cham; 2024. p. 47–64.
40. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. In: 3rd International Conference on Learning Representations (ICLR 2015). Computational and Biological Learning Society, ???; 2015. p. 1–14.
41. Abbasi S, Bilal M, Mukherjee T, Churm J, Pournik O, Epiphaniou G, Arvanitis T. Deep learning-based synthetic skin lesion image classification. Stud Health Technol Inf. 2024;316:1145–50.
42. Tognetti L, Cartocci A, Cinotti E, Moscarella E, Farnetani F, Carrera C, Lallas A, Tiodorovic D, Longo C, Puig S, Perrot J, Argenziano G, Pellacani G, Gennaro C, Balistreri A, Cevenini G, Rubegni P. Dermoscopy of early melanomas: variation according to the anatomic site. Arch Dermatol Res. 2022;314:183–90.
43. Tschandl P. The HAM10000 Dataset, a Large Collection of Multi-source Dermatoscopic Images of Common Pigmented Skin Lesions. https://doi.org/10.7910/DVN/DBW86T
44. Rotemberg V, Kurtansky N, Betz-Stablein B, Caffery L, Chousakos E, Codella N, Combalia M, Dusza S, Guitera P, Gutman D, Halpern A, Helba B, Kittler H, Kose K, Langer S, Lioprys K, Malvehy J, Musthaq S, Nanda J, Soyer P. A patient-centric dataset of images

and metadata for identifying melanomas using clinical context. Scientific Data. 2021;8:34.
45. Tognetti L, Cevenini G, Moscarella E, Cinotti E, Farnetani F, Mahlvey J, Perrot JL, Longo C, Pellacani G, Argenziano G, Fimiani M, Rubegni P. An integrated clinical-dermoscopic risk scoring system for the differentiation between early melanoma and atypical nevi: the idscore. J Eur Acad Dermatol Venereol. 2018;32(12):2162–70.
46. Luschi A. DCGAN. 2025. https://github.com/alessioluschi/DCGAN. Accessed 08 Mar 2025
47. Karras T, Aittala M, Laine S, Härkönen E, Hellsten J, Lehtinen J, Aila T. Alias-free generative adversarial networks. In: Proceedings of the 35th International Conference on Neural Information Processing Systems. NIPS '21. Curran Associates Inc., Red Hook, NY, USA; 2021.
48. Tognetti L, Cartocci A, Żychowska M, Savarese I, Cinotti E, Pizzichetta MA, Moscarella E, Longo C, Farnetani F, Guida S, Paoli J, Lallas A, Tiodorovic D, Stanganelli I, Magi S, Dika E, Zalaudek I, Suppa M, Argenziano G, Pellacani G, Perrot JL, Miracapillo C, Rubegni G, Cevenini G, Rubegni P. A risk-scoring model for the differential diagnosis of lentigo maligna and other atypical pigmented facial lesions of the face: The facial idscore. J Eur Acad Dermatol Venereol. 2023;37(11):2301–10.
49. Luschi A. Dermatological Syntethic Image Assessment (DSIA) platform. 2025. https://gan.idscore.net/. Accessed 08 Mar 2025

Authors and Affiliations

Alessio Luschi[1] · Linda Tognetti[2] · Alessandra Cartocci[2] · Gabriele Cevenini[1] · Pietro Rubegni[2] · Ernesto Iadanza[1,3]

✉ Alessio Luschi
alessio.luschi@unisi.it

Linda Tognetti
linda.tognetti@dbm.unisi.it

Alessandra Cartocci
alessandra.cartocci@dbm.unisi.it

Gabriele Cevenini
gabriele.cevenini@unisi.it

Pietro Rubegni
pietro.rubegni@unisi.it

Ernesto Iadanza
ernesto.iadanza@unisi.it

[1] Department of Medical Biotechnologies, University of Siena, Via Aldo Moro 2, 53100 Siena, SI, Italy

[2] Dermatology Unit, Department of Medicine, Surgery and Neuroscience, University of Siena, Strade delle Scotte 14, 53100 Siena, SI, Italy

[3] Council of Societies, International Federation for Medical and Biological Engineering, 5, Rue Auguste Vacquerie, 75116 Paris, France