ISSN 2282-0930 • Epidemiology Biostatistics and Public Health - XIII[th] SISMEC Congress - Vol. 20 Suppl. 1

Real or Synthetic? Dermatologist Agreement on Synthetic *vs.* Real Melanoma and Pattern Recognition

Cartocci Alessandra[(1)], Luschi Alessio[(2)], Tognetti Linda[(1)], Iadanza Ernesto[(2)], Rubegni Pietro[(1)]

(1) Dermatology Unit, Department of Medicine, Surgery and Neuroscience, University of Siena, Italy
(2) Department of Medical Biotechnologies, University of Siena, Italy

CORRESPONDING AUTHOR: *Cartocci Alessandra, alessandra.cartocci@unisi.it*

BACKGROUND

The validation of synthetic dermatological images generated by Generative Adversarial Networks (GANs) [1] is crucial for their integration into clinical and research workflows. Despite rapid progress in image synthesis, a standardized framework for evaluating the realism and diagnostic utility of synthetic skin lesions through expert review is still lacking [2]. Existing automated evaluation metrics, while informative, do not always align with human perception and diagnostic expectations. Particularly in medical domains, subtle visual cues and contextual interpretation often elude algorithmic assessment [3]. Human evaluations remain the most direct means of determining whether synthetic images capture the nuanced features necessary for clinical utility. Without structured expert-based validation, synthetic images may introduce bias or mislead models and clinicians, hampering their responsible deployment in diagnostic support systems, training datasets, or educational tools.

OBJECTIVES

This study aims to conduct an expert-based qualitative evaluation of synthetic melanoma images. Specifically, it investigates the subjective perception of image realism, diagnostic quality, and the recognizability of key dermoscopic features. By engaging dermatologists in a blinded assessment of synthetic and real images, we seek to establish a foundation for systematically validating synthetic dermatological data for use in AI development, medical education, and clinical decision support. This work emphasizes the importance of subjective expert validation as a complement to technical performance metrics in assessing the fidelity of GAN-generated skin lesion images.

MATERIALS AND METHODS

StyleGAN3-T [4] was trained on a dataset of dermoscopic images of melanoma [5–7] with adaptive discriminator augmentation and transfer learning. A total of 25 synthetic melanoma images were generated and randomly mixed with 25 real melanoma images, resulting in a 50-image dataset. Seventeen board-certified dermatologists with varying levels of experience (low <4 years, medium 5–8 years, high >8 years) participated in the evaluation. Participants were blinded to image origin and asked to classify each image as real or synthetic. They also assessed the presence of 16 defined dermoscopic patterns according to standardized definitions and rated four dimensions—image quality, skin texture, visual realism, and color realism—on a 7-point Likert scale. Additionally, participants reported their confidence in each classification decision. Statistical analyses included Chi-square tests for categorical comparisons, and Fleiss' Kappa and Krippendorff's Alpha were used to measure inter-rater agreement.

DOI: 10.54103/2282-0930/29361

RESULTS

Real images were consistently rated higher than synthetic images across all qualitative dimensions: image quality (high: 15.8% real vs. 11.3% synthetic), skin texture (high: 22.4% vs. 13.4%), and visual realism (high: 22.6% vs. 13.2%), all with $p < 0.001$. Confidence in evaluations was also significantly greater for real images, with high confidence reported in 17.4% of real cases compared to 8.7% for synthetic ones ($p < 0.001$). Regarding the recognition of image origin, the overall classification accuracy was 64%. Real images were correctly identified in 73% of cases, while only 56% of synthetic images were correctly classified as synthetic. Accuracy increased with expertise: from 59% in the low-experience group to 71% among high-experience dermatologists. Similarly, higher self-reported confidence was associated with improved performance (accuracy 74% at high confidence level).

Recognition of specific dermoscopic features showed differences between real and synthetic images. The blue-white veil was detected in 29.1% of real images compared to 13.8% of synthetic ones ($p < 0.001$), and shiny white streaks in 22.6% vs. 7.9% ($p < 0.001$). Conversely, synthetic images were more frequently associated with irregular pigmented blotches (45.0% vs. 30.9%, $p < 0.001$). The multicomponent pattern, typically indicative of melanoma complexity, was identified in 40.6% of real images versus only 23.2% of synthetic ones ($p < 0.001$), suggesting a gap in the synthetic images' structural fidelity (Table 1).

Inter-rater agreement for the classification of real versus synthetic images was low, with a Fleiss' kappa of 0.183. Pattern recognition agreement also remained weak (e.g., kappa < 0.3 for most features), underscoring variability in expert interpretations. Further subgroup analyses showed that images rated as highly realistic or evaluated with high confidence were more likely to be classified correctly, with accuracy rising to 74% in the highest-confidence subgroup.

CONCLUSIONS

Synthetic melanoma lesions generated using StyleGAN3-T demonstrate visually convincing features and were frequently perceived as real, yet consistently underperformed compared to real images in diagnostic quality and structural detail. Participants often struggled to distinguish synthetic from real lesions, particularly when realism ratings were medium to high. Critical diagnostic patterns, such as the blue-white veil and shiny white streaks, were significantly underrepresented in synthetic images. These limitations were reflected in the lower classification confidence and weaker inter-rater agreement.

Despite these challenges, the study highlights the potential of synthetic data to approach realism levels sufficient for research and educational use. Qualitative validation by dermatologists is essential to benchmark the readiness of synthetic images for real-world medical applications. As generative models continue to evolve, expert evaluation should remain a key component of validation pipelines to ensure clinical and pedagogical reliability.

REFERENCES

1. Goodfellow, I., Pouget-Abadie, J., Mirza, M., et al., Generative adversarial networks. Commun. ACM 63(11), 139–144 (2020)
2. M. Kang, J. Shin, and J. Park, "Studiogan: A taxonomy and benchmark of gans for image synthesis," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 45, no. 12, pp. 15 725–15 742, 2023.
3. D. A. Chan and S. P. Sithungu, "Evaluating the suitability of inception score and frechet inception distance as metrics for quality and diversity in image generation," in Proceedings of the 2024 7th International Conference on Computational Intelligence and Intelligent Systems, ser. CIIS '24. Association for Computing Machinery, 2025, p. 79–85.
4. Karras, T., Laine, S., Aila, T., A Style-Based Generator Architecture for Generative Adversarial Networks. IEEE Transactions on Pattern Analysis & Machine Intelligence 43(12), 4217–4228 (2021)
5. 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
6. Rotemberg, V., Kurtansky, N., Betz-Stablein, B., et al., .: A patient-centric dataset of images and metadata for identifying melanomas using clinical context. Scientific Data 8, 34 (2021)
7. Tognetti, L., Cevenini, G., Moscarella, et al., An integrated clinical-dermoscopic risk scoring system for the differentiation between early melanoma and atypical nevi: the id-score. Journal of the European Academy of Dermatology and Venereology 32(12), 2162–2170 (2018)

DOI: 10.54103/2282-0930/29361

Table 1 - Distribution of dermatological patterns identified by edrmatologists during the assessment between real and synthetic. P-values reflect the statistical significance of differences between presence of patterns in the two classes. Results are grouped by level of expertise. The level of agreement between raters for the two classes is also reported (Fleiss' Kappa)

Pattern	Real	Synthetic	p-value	Fleiss' Kappa (real)	Fleiss' Kappa (synthetic)
	N=340	N=340			
Atypical network	241 (70.9)	245 (72.1)	0.799	0.175	0.151
Hypopigmented areas	97 (28.5)	65 (19.1)	0.005	0.205	0.149
Irregular dots and globules	126 (37.1)	99 (29.1)	0.034	0.198	0.204
Irregular streaks	59 (17.4)	48 (14.1)	0.292	0.149	0.124
Irregular pigmented blotches	105 (30.9)	153 (45.0)	<0.001	0.261	0.34
Blue with veil	99 (29.1)	47 (13.8)	<0.001	0.336	0.34
Blue-grey globules	29 (8.5)	8 (2.4)	0.001	0.053	0.008
Blue-grey peppering	20 (5.9)	4 (1.2)	0.002	0.077	0.020
White scar-like areas	71 (20.9)	39 (11.5)	0.001	0.268	0.236
Shiny white streaks	77 (22.6)	27 (7.9)	<0.001	0.381	0.381
Atypical vascular pattern	20 (5.9)	11 (3.2)	0.141	0.223	0.213
Pink areas	85 (25.0)	58 (17.1)	0.014	0.245	0.189
Reticular pattern	59 (17.4)	63 (18.5)	0.764	0.067	0.014
Globular pattern	16 (4.7)	4 (1.2)	0.013	0.197	0.083
Homogenous pattern	30 (8.8)	29 (8.5)	1.000	0.086	0.039
Multicomponent pattern	138 (40.6)	79 (23.2)	<0.001	0.114	0.087

DOI: 10.54103/2282-0930/29361