Diagnostic Concordance and Morphological Realism of Synthetic Dermatological Images Generated by NANO BANANA 2

An Exploratory Study

Jonathan Shapiro[1], Yuliya Valdman[2], Ibrahim Alatawna[2], Adi Saadia2, Lyakhovitsky Anna[3], Mor Atlas[4], Romi Gleicher5, Yotam Adi[2]

1. Maccabi Healthcare Services 2. Department of Dermatology and Venereology, Soroka Medical Center, Beer Sheva, Israel 3. Department of Dermatology, Sheba Medical Center Ramat-Gan Israel 4. Ono Academic Collage, Israel 5. Bruce Rappaport Faculty of Medicine, Technion - Institute of Technology, Israel, Dermatology Department

Dermatologists' overall diagnostic concordance	AI models overall diagnostic concordance	of synthetic images classified as real by AI
83.3%	**89.3%**	**91.9%**
across 3 dermatologists, 95% CI 78.4–87.3%	across 3 AI models, 95% CI 85.0–92.4%	mean across 3 AI models, ± 6.5%

Objective

- The scarcity of clinical images for rare skin diseases may limit the development of AI for these specific conditions [1].
- While generative AI for creating synthetic images offers a potential solution [2–3], prior tools achieved a diagnostic concordance of only up to 40% [4].
- To evaluate the diagnostic viability of NANO BANANA 2 synthetic images using board-certified dermatologists and AI models, helping to identify whether diagnostic failures stem from image-quality limitations, rare disease complexity, or nonspecific visual features.

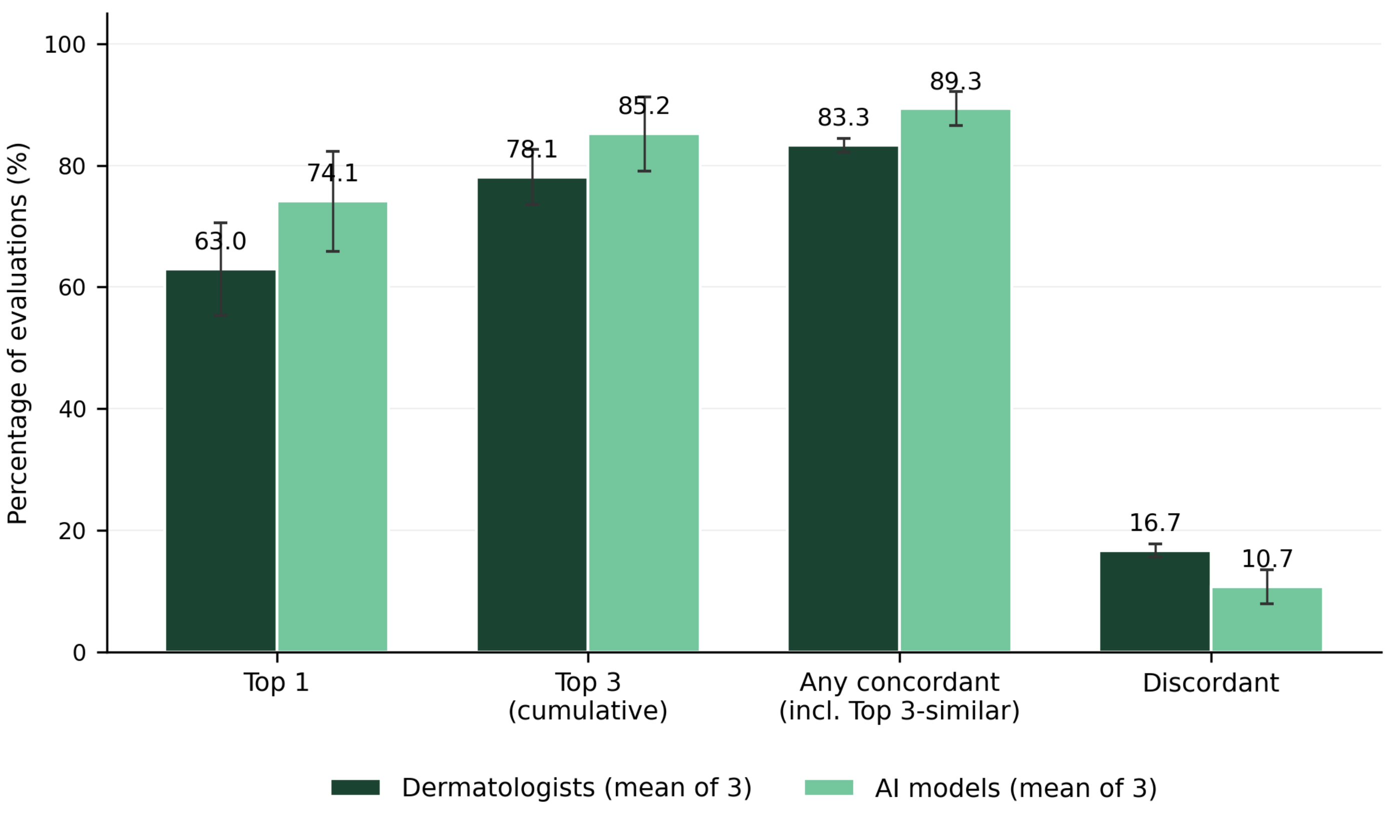

Fig 1. Diagnostic concordance distribution: dermatologists vs. AI models.

Methods

Dataset: We used the NANO BANANA 2 model to generate 90 synthetic clinical images representing various dermatologic conditions. The generation prompt served as the ground-truth diagnosis.

Evaluators: Images were independently reviewed by 3 board-certified dermatologists and 3 multimodal AI models (Claude, ChatGPT, Gemini), yielding 540 total evaluations.

Diagnostic concordance: Raters provided a differential diagnosis of three conditions, ranked by probability. Diagnoses were classified as concordant if they fell into one of three categories: **Top 1** (primary diagnosis matched the ground-truth), **Top 3** (ground-truth included within the three-disease differential), or **Top 3-similar** (a morphologically related diagnosis appeared in the differential). Otherwise, the diagnosis was classified as **Discordant**.

Recovery step: To distinguish synthetic image-quality failures from inherent disease difficulty, discordant cases underwent a retrospective review. Both dermatologists and AI models were shown a real clinical photograph corresponding to the same ground-truth prompt and asked for a revised diagnosis, scored using the same concordance scheme.

Quality ratings: Dermatologists rated image realism on a 1–5 scale (1 = obvious artifact, 5 = highly realistic), while AI models classified each image binary as real or synthetic.

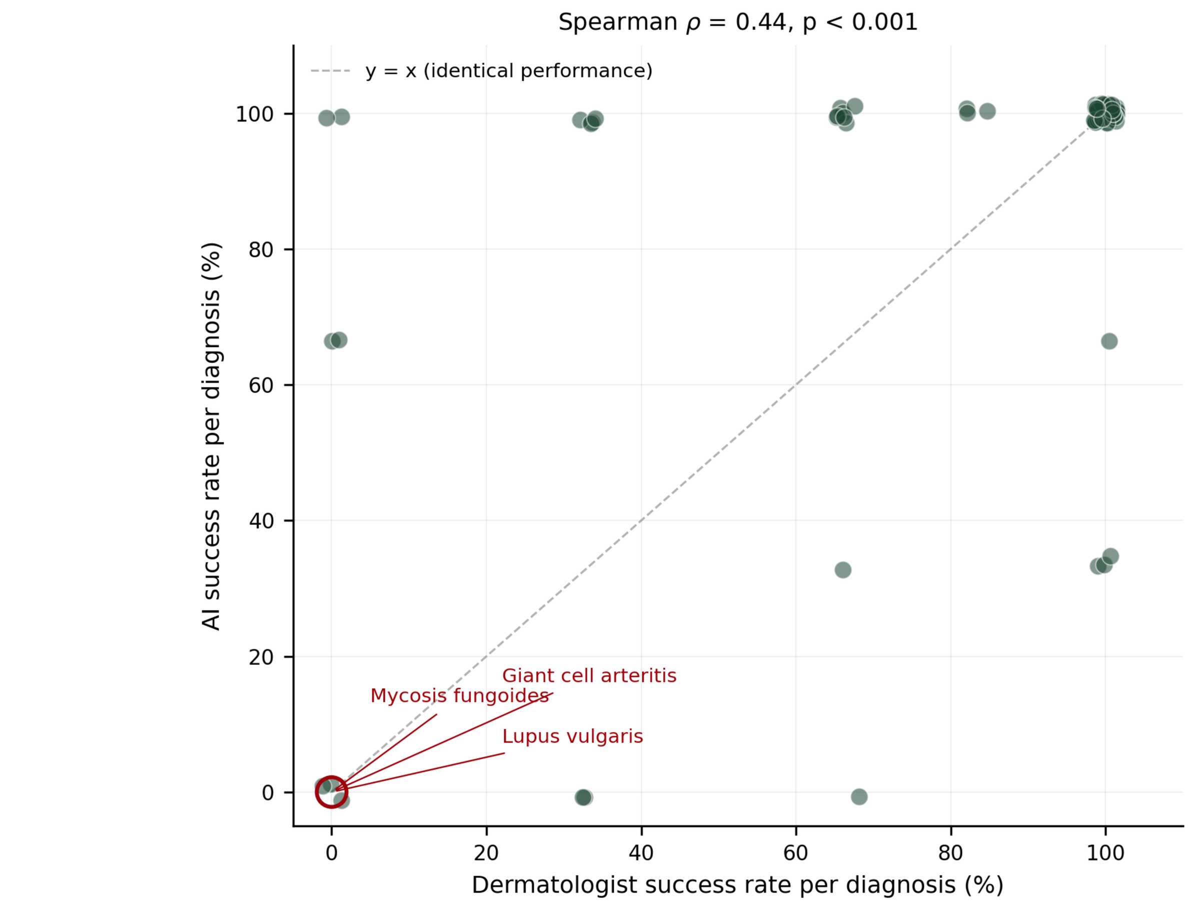

Fig 2. Per-diagnosis concordance: AI vs. dermatologists. Each dot is one disease (n = 80).

Results

- Diagnostic concordance for both groups is presented in **Figure 1**, with no significant difference found between them (p = 0.061).
- **Figure 2** details per-diagnosis concordance; notably, three conditions were discordant across both groups (giant cell arteritis, lupus vulgaris, and mycosis fungoides).
- **Figure 3** presents recovery rates on real images, which were significantly higher for dermatologists than for AI models (84.4% vs. 41.4%).
- The evaluation of image realism by AI models and dermatologists is presented in **Figures 4 and 5**, respectively: AI classified 91.9% of the synthetic images as real, while dermatologists rated 55.5% as 'Realistic' or 'Highly Realistic'.
- The analysis of inter-rater agreement is presented in **Table 1**, demonstrating that AI models achieved higher agreement than the board-certified dermatologists.

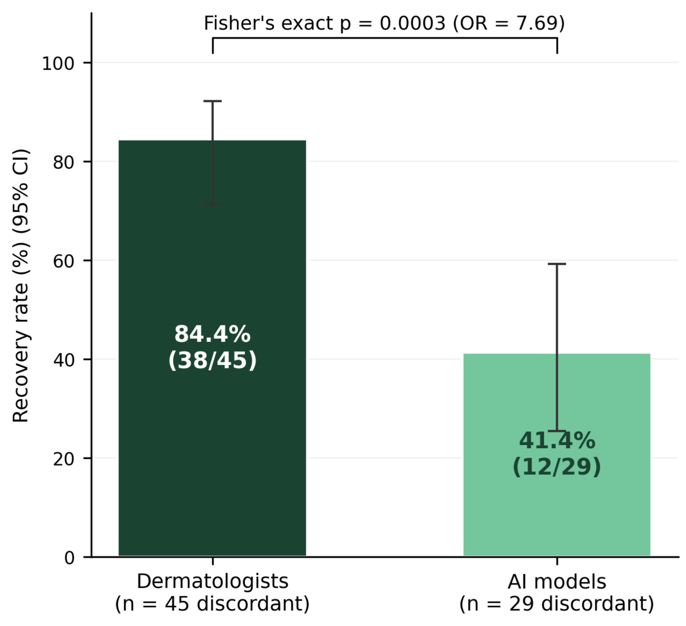

Fig 3. Recovery on real image (derm n = 45, AI n = 29).

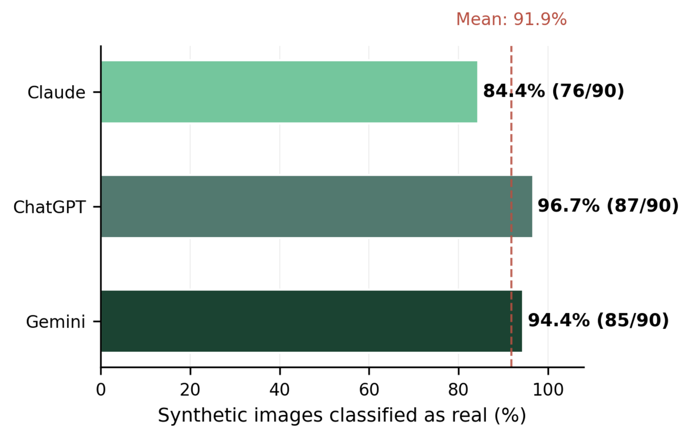

Fig 4. Synthetic images classified as real by AI (per model).

Table 1. Performance comparison: dermatologists vs. AI models.

Metric	Dermatologists	AI Models
Top 1 (mean ± SD)	63.0% ± 7.6%	74.1% ± 8.2%
Overall Concordance	83.3%	89.3%
Majority Concordance (image)	85.6%	88.9%
Recovery on real image	84.4%	41.4%
Fleiss' κ (binary)	0.52	0.73

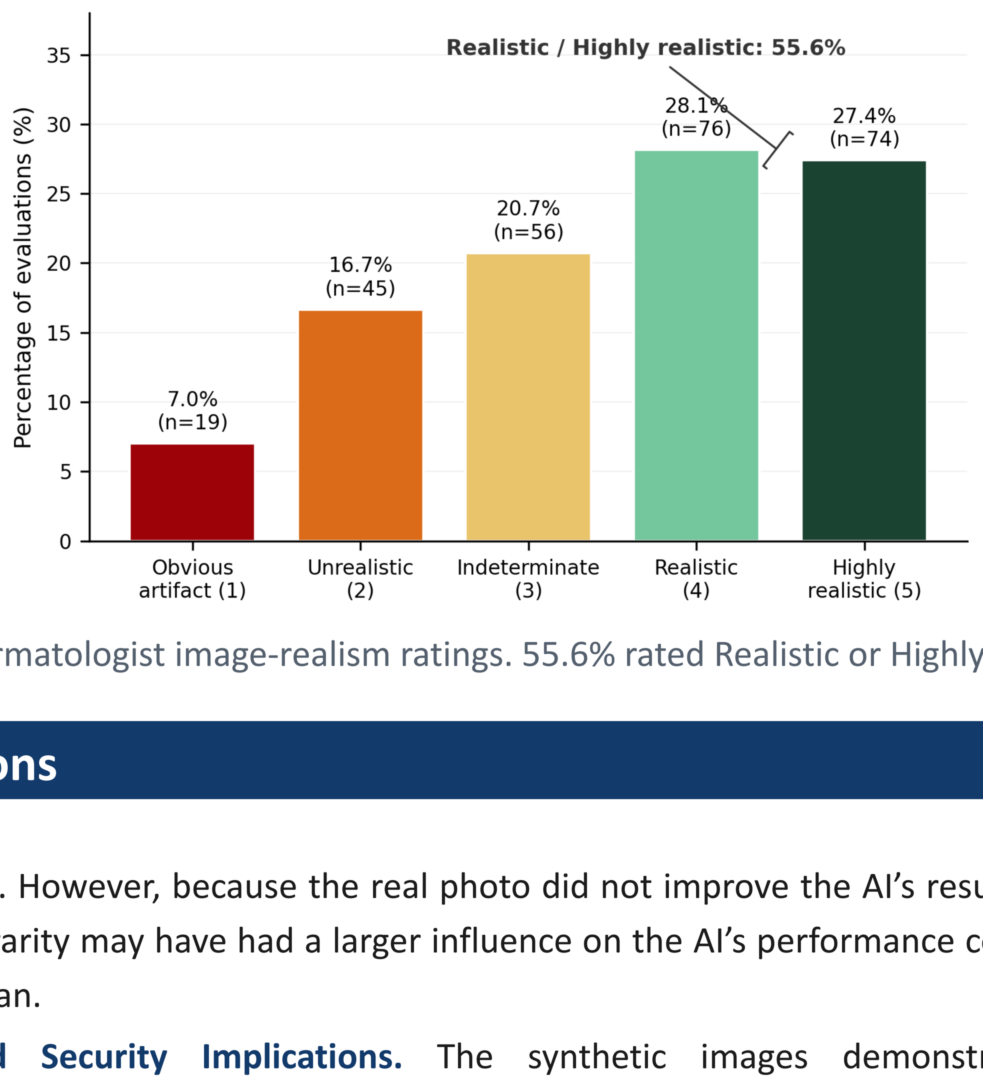

Fig 5. Dermatologist image-realism ratings. 55.6% rated Realistic or Highly realistic.

Discussion and Conclusions

Improved Baseline Concordance. Overall diagnostic concordance exceeded rates reported in prior studies utilizing alternative generative AI tools, with no significant difference observed between AI and dermatologist performance.

Shared Diagnostic Vulnerabilities. Dermatologists and AI models failed on a highly correlated subset of conditions, indicating shared clinical blind spots ($\rho = 0.44$, $p < 0.001$).

Mechanisms of Diagnostic Failure. The significant discrepancy in recovery rates (84.4% for dermatologists vs. 41.4% for AI) may point toward different underlying causes for the discordant diagnoses. For dermatologists, the ability to easily establish a concordant diagnosis upon seeing the real photo suggests the initial error stemmed from the synthetic image quality. However, because the real photo did not improve the AI's results, it implies that disease rarity may have had a larger influence on the AI's performance compared to a human clinician.

Realism and Security Implications. The synthetic images demonstrated strong morphological realism across both evaluator groups. These results suggest that NANO BANANA 2 has the potential to produce images of high visual quality that are convincing to both automated systems and expert human raters. However, the significant discrepancy in classification rates between the AI models and the dermatologists supports the hypothesis that AI cannot yet serve as a reliable tool for distinguishing real from AI-generated clinical images, raising concerns about the potential for misuse of synthetic clinical photographs.

References.

1. Choy SP, et al. Systematic review of deep learning image analyses for the diagnosis and monitoring of skin disease. NPJ Digital Medicine. 2023;6(1):180. 2. Kim M, Lee J, Park S. Diffusion-based skin disease data augmentation with fine-grained detail preservation and interpolation for data diversity. PLOS ONE. 2025. doi:10.1371/journal.pone.0319673. 3. Wang J, Wang K, Yu Y, et al. Self-improving generative foundation model for synthetic medical image generation and clinical applications. Nature Medicine. 2024. doi:10.1038/s41591-024-03359-y. 4. Kooper-Johnson S, Lim S, Aldhalaan J, et al. Image Generation of Common Dermatological Diagnoses by Artificial Intelligence: Evaluation Study of the Potential for Education and Training Purposes. JMIR Dermatology. 2025;8:e72371. doi:10.2196/72371.

Contact: jonmidi@gmail.com • AI in Dermatology Summit 2026 — Vienna, Austria