iFSCC 35TH
CONGRESS
FRANCE_CANNES 15-18 SEPT 25
THE FUTURE IS SCIENCE
ifscc2025.com
IFSCC
SOCIÉTÉ FRANÇAISE DE COSMÉTOLOGIE

IFSCC 2025 full paper (IFSCC2025-1175)

“Validation of a Machine Learning-Based Model for Blackhead Evaluation in a Pilot Trials”

Guansheng Yang [1,*], Hua Sun [1], Gang Jin [1] and Yanwen Jiang [1]

[1] Shanghai China-norm Quality Technical Service Co., Ltd., Shanghai, China 1

1. Introduction

Blackheads (open comedones) remain a common cosmetic concern and an early clinical sign of acne vulgaris, is a very common human skin disease that affects up to 85 % of adolescents and up to 50% adults [1,2]. Traditional assessments rely on manual lesion counts or photographic scales, techniques that are labor-intensive and prone to inter-observer variability. Recent surveys report inter-grader intraclass correlation coefficients (ICCs) as low as 0.45 for dermatological image segmentation, highlighting substantial subjectivity [3].

Computer-vision tools for acne and comedone analysis have advanced quickly in the past two years. A 2025 study on AcneDGNet reported a mean counting error of only 1.9 ± 3.3 lesions while keeping whole-face grading accuracy close to 90 % across VISIA, DSLR and phone images [4]. Similar progress appears elsewhere. AcneDet, built on Faster R-CNN, detects blackheads and other lesion types in smartphone photos and keeps F1 scores above 0.90 on an external test set [5]. A lightweight clinic-app reached acne sensitivity of 89 % and specificity of 93 %, values already near dermatologist performance [6]. Segmentation models have improved too: the hybrid encoder HDS-Net now posts Dice coefficients > 0.91 on public skin-lesion benchmarks [7].

Taken together, recent studies show that automated intelligent model can match expert graders under varied imaging conditions, making a prospective clinical test next step.

The present pilot trial validates our deep-learning model for nasal blackhead quantification in a controlled clinical setting. The goal is to show that, with standardized VISIA-CR photographs, the model can equal expert accuracy and at the same time lower inter-grader variation. A YOLO-based machine learning model trained on nose images has already reached overall accuracy 0.9 and ICC values above 0.9 across all severity groups, proving that expert-level agreement is possible on this challenging area [8]. We therefore compared model outputs with expert counts using Two-way mixed effects intraclass correlation coefficients for absolute agreement and Bland–Altman plots for bias. We expected good reliability and clinically trivial limits of agreement, outcomes that would support the model’s use in efficacy studies and personalized skin-care assessments.

2. Materials and Methods

2.1 Participants

A total of 100 healthy volunteers presenting visible blackheads on the nose were enrolled after written consent.

2.2 Imaging

Photographs were taken with a Canfield VISIA-CR under its built-in chin rest and forehead pads to keep head position fixed. Three angles were recorded - front, left and right - using standard lighting.

Before imaging, subjects washed the face with tap water and acclimated for 30 min to let the skin reach room conditions (21± 1 °C, 50 ± 10% RH). No make-up or skincare was allowed during that time.

2.3 Expert grading

One validated expert grader, blinded to the model outputs, counted open comedones by reviewing the photographs in a darkened room on a color-calibrated monitor; to limit visual fatigue and preserve grading reliability, each scoring session was kept to no more than 1 hour.

2.4 Machine-learning model

The automated counts were produced with the nasal-blackhead model described by Sun et al. 2024 [8]. 350 VISIA nasal photos were collected from adults (18-60 y) and hand-marked each blackhead by bounding box. The images were then used to fine-tune a lightweight YOLO-v5 detector, chosen for speed and small-object sensitivity, reached algorithm accuracy above 0.9.

2.5 Statistical analysis

Model versus expert agreement was tested with a two-way mixed-effects for absolute agreement; values ≥ 0.90 were classed as excellent, 0.75–0.89 good, < 0.75 poor, as suggested by Shrout & Fleiss and later reliability reviews [9,10]. Bias was visualized with Bland-Altman plots and 95 % limits of agreement (mean difference ± 1.96 SD) [11].

3. Results

3.1 Participant Characteristics

Table 1 shows that the expert blackhead count was 81.91 ± 4.72 (mean ± SE), whereas the model returned 82.49 ± 5.52 (mean ± SE). A paired-samples t-test found no significant difference between the two means (p = 0.89), indicating statistical equivalence in average performance.

Table 1. Blackheads on the nose as determined by expert assessment and model counting

	N	Expert counting	Model counting	P value
Number of blackheads on the nose	100	81.91±4.72	82.49±5.52	0.890

Data showed as Mean±SE

3.2 Reliability Metrics

The results of the consistency analysis of the expert assessment and model counting are shown in Table 2. The intraclass correlation coefficient for the number of nasal blackheads is 0.8 (p<0.05) with good consistency.

Table 2. Consistency analysis of expert assessment and model counting

	N	ICC	P
Number of blackheads on the nose	100	0.801	<0.001

Data showed as Mean±SE

3.3 Agreement Analysis

Chart 1 gives the Bland–Altman plot. The mean difference (model – expert) was +0.58 comedones, signifying a slight, clinically trivial over-count by the algorithm. The 95 % limits of agreement were -82.87 to 81.71 comedones. Only 7 of 100 pairs (7 %) lay outside these limits, a proportion that closely matches the statistical expectation that roughly 5 % of points will fall beyond the 95 % LoA when two methods agree well, thereby confirming that the automated counts are interchangeable with expert counts for practical use [13].

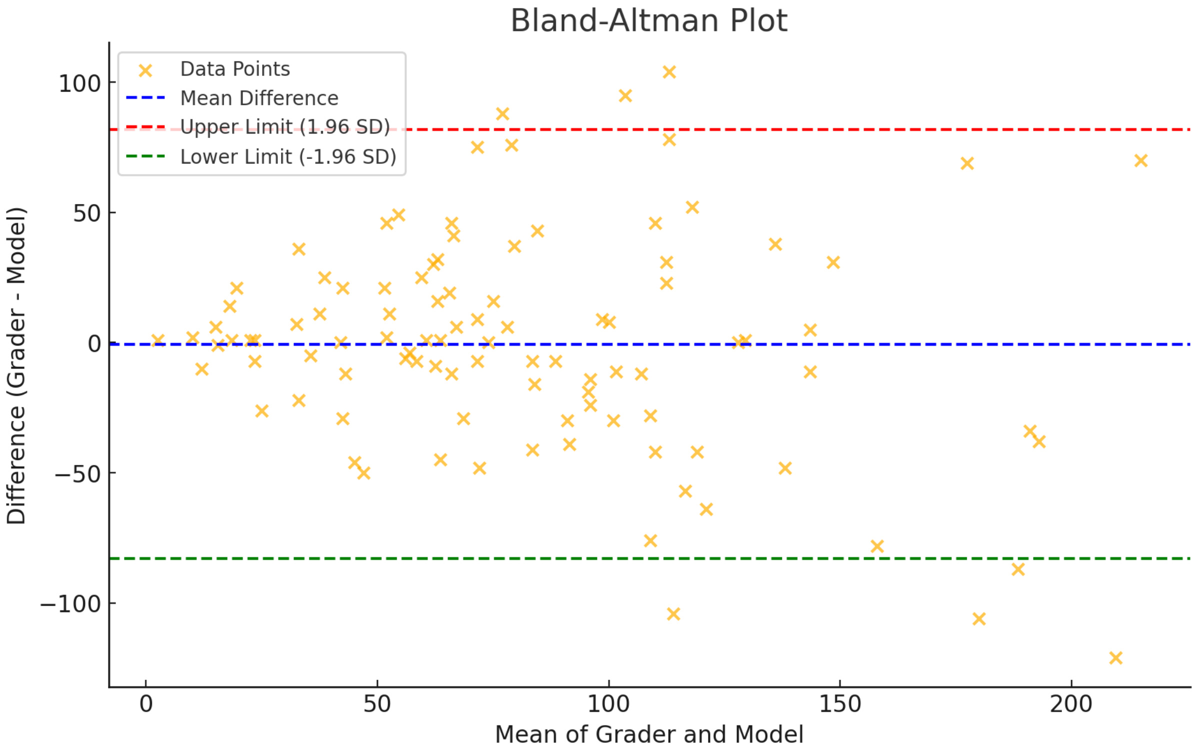

Chart 1. Bland–Altman plot of expert grader versus model counts. The horizontal solid line marks the mean bias (+0.58), while the dashed lines indicate the 95 % limits of agreement (−82.87 to +81.71). Each point represents one of 100 nasal photographs.

4. Discussion

The agreement metrics confirm that our deep-learning model can stand in for human graders in routine blackhead studies, the proportion statistically expected when two methods accord well. Automation also reduced reading time from about 10 minutes to 1 seconds per image, mirroring the efficiency gains reported for other semi-automatic dermatology tools [14].

These findings place the model at the performance range of recent acne-detection networks (mAP 0.54–0.89) and remote eczema-scoring systems (ICC ≈ 0.74–0.94) [5,15].

In this pilot trail, all photographs were captured with a VISIA-CR system that fixes head position and uses controlled multi-modal lighting; this hardware has been shown to keep test–retest error below 3 % for facial features [16]. While ICC and Bland–Altman limits were used to give a fuller picture of relative and absolute agreement. Our sample of 100 is moderate rather than large, it is still larger than the median of 65 seen in published agreement studies and exceeds the 50–60 subjects usually suggested to obtain ICC precision of ±0.10. This size therefore provides a reasonable balance between logistical feasibility and statistical stability without overstating the dataset's breadth [17].

For cosmetic-efficacy trials, automated counting reduces subjectivity, enabling smaller sample sizes or tighter non-inferiority margins. In personalized skincare, near-real-time feedback could guide consumer product selection and monitor treatment adherence.

Our cohort came from a single research centre, so ethnic representation is limited. The study focused exclusively on the nose, yet acne commonly affects cheeks and forehead. In addition, counting accuracy is sensitive to image quality—blur, poor focus or exposure can still mislead the algorithm.

Future work should therefore widen the training set and test conditions: (i) collect images from multiple sites and skin-tone groups; (ii) add full-face photographs so the network can recognise high-incidence areas such as the forehead while maintaining current accuracy; and (iii) incorporate automatic color-calibration and blur-detection steps, allowing the system to move from controlled VISIA photography to everyday smartphone pictures without loss of reliability.

To enrich biological insight, multimodal signals could also be integrated—for example, porphyrin fluorescence or UV-excited sebaceous imaging—to link comedone counts with microbial and sebaceous activity, a direction already explored in recent fluorescence studies [18].

By addressing these points, the model can evolve from a nasal-specific counter into a robust, full-face, multi-centre platform that supports both clinical trials and personalized skincare recommendations.

5. Conclusion

This study confirms the reliability of a machine learning-based blackhead evaluation model, demonstrating high consistency with expert evaluations. The model effectively reduces subjectivity, improves efficiency, and offers a practical alternative to traditional methods. Its scalability and objective approach position it as a transformative tool in cosmetic testing and dermatological research. This model has the potential to set a new benchmark for automated skin condition assessment and contribute to advancements in personalized skincare and AI-driven clinical diagnostics.

Reference

1. Ramli, R., Malik, A.S., Hani, A.F.M. and Jamil, A. (2012), Acne analysis, grading and computational assessment methods: an overview. Skin Research and Technology, 18: 1-14. https://doi.org/10.1111/j.1600-0846.2011.00542.x
2. Tan, J. K. (2008). Current measures for the evaluation of acne severity. Expert Review of Dermatology, 3(5), 595–603. doi:10.1586/17469872.3.5.595
3. Hurault G, Pan K, Mokhtari R, et al. Detecting Eczema Areas in Digital Images: An Impossible Task?. JID Innov. 2022;2(5):100133. Published 2022 May 23. doi:10.1016/j.xjidi.2022.100133
4. Gao, N., Wang, J., Zhao, Z. et al. Evaluation of an acne lesion detection and severity grading model for Chinese population in online and offline healthcare scenarios. Sci Rep 15, 1119 (2025). https://doi.org/10.1038/s41598-024-84670-z
5. Huynh QT, Nguyen PH, Le HX, Ngo LT, Trinh N-T, Tran MT-T, Nguyen HT, Vu NT, Nguyen AT, Suda K, et al. Automatic Acne Object Detection and Acne Severity Grading Using Smartphone Images and Artificial Intelligence. Diagnostics. 2022; 12(8):1879. https://doi.org/10.3390/diagnostics12081879
6. Wang J, Luo Y, Wang Z, et al. A cell phone app for facial acne severity assessment. Appl Intell (Dordr). 2023;53(7):7614-7633. doi:10.1007/s10489-022-03774-z
7. Xue Y, Chen X, Liu P, Lv X. HDS-Net: Achieving fine-grained skin lesion segmentation using hybrid encoding and dynamic sparse attention. PLoS One. 2024;19(3):e0299392. Published 2024 Mar 21. doi:10.1371/journal.pone.0299392
8. Sun H, Yang G, Yuan J, Jiang Y, Jin G. Establishment and verification of a method for analyzing nasal blackheads images. Skin Res Technol. 2024;30(3):e13648. doi:10.1111/srt.13648
9. Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. Psychol Bull. 1979;86(2):420-428. doi:10.1037//0033-2909.86.2.420
10. Koo TK, Li MY. A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research [published correction appears in J Chiropr Med. 2017 Dec;16(4):346. doi: 10.1016/j.jcm.2017.10.001.]. J Chiropr Med. 2016;15(2):155-163. doi:10.1016/j.jcm.2016.02.012
11. Giavarina D. Understanding Bland Altman analysis. Biochem Med (Zagreb). 2015;25(2):141-151. Published 2015 Jun 5. doi:10.11613/BM.2015.015
12. Sedgwick P. Limits of agreement (Bland-Altman method) BMJ 2013; 346 :f1630 doi:10.1136/bmj.f1630
13. Zhang W, Zeng J, Huang Q, Liu Z, Li J. The feasibility analysis of calculating proptosis by simple Pythagorean theorem. Eur J Ophthalmol. 2021;31(2):397-404. doi:10.1177/1120672120901704
14. Gronenschild, E. H. B. M., Muris, D. M. J., Schram, M. T., Karaca, Ü., Stehouwer, C. D. A., & Houben, A. J. H. M. (2013). Semi-automatic assessment of skin capillary density: Proof of principle and validation. Microvascular Research, 90, 192–198. doi:10.1016/j.mvr.2013.08.003
15. Ragamin A, Schappin R, Tan Nguyen N, et al. Remote severity assessment in atopic dermatitis: Validity and reliability of the remote Eczema Area and Severity Index and Self-Administered Eczema Area and Severity Index. JAAD Int. 2023;13:184-191. Published 2023 Aug 28. doi:10.1016/j.jdin.2023.07.019
16. Henseler H. Assessment of the reproducibility and accuracy of the Visia® Complexion Analysis Camera System for objective skin analysis of facial wrinkles and skin age. GMS Interdiscip Plast Reconstr Surg DGPW. 2023;12:Doc07. Published 2023 Oct 2. doi:10.3205/iprs000177

17. Han O, Tan HW, Julious S, et al. A descriptive study of samples sizes used in agreement studies published in the PubMed repository. BMC Med Res Methodol. 2022;22(1):242. Published 2022 Sep 19. doi:10.1186/s12874-022-01723-5
18. Chekanov K, Danko D, Tlyachev T, Kiselev K, Hagens R, Georgievskaya A. State-of-the-Art in Skin Fluorescent Photography for Cosmetic and Skincare Research: From Molecular Spectra to AI Image Analysis. Life (Basel). 2024;14(10):1271. Published 2024 Oct 6. doi:10.3390/life14101271