Dermatology Reports
https://journals.pagepress.net/dr

eISSN 2036-7406

Please cite this article as:

Shkodrani E, Xhaja A, Shameti M, Shkodrani B. Comparison between physician assessment and artificial intelligence analysis in digital dermoscopy: evaluation of 71 dermoscopic cases. *Dermatol Rep 2026 [Epub Ahead of Print] doi: 10.4081/dr.2026.10872*

Received: 15 Marc 2025; Accepted: 14 July 2026.

Comparison between physician assessment and artificial intelligence analysis in digital dermoscopy: evaluation of 71 dermoscopic cases

Entela Shkodrani,[1] Alert Xhaja,[1] Martin Shameti,[2] Barbara Shkodrani[3]

[1]Clinic of Dermatology, University Hospital Center "Mother Theresa", Tirana, Albania; [2]Polytechnic University of Tirana, Albania; [3]Department of Preventive Biomedicine, University of Rome Tor Vergata, Rome, Italy

Correspondence: Entela Shkodrani, Clinic of Dermatology, University Hospital Center "Mother Theresa", Rr. Dibra, n. 362, Tirana, Albania. E-mail: entela.shkodrani@gmail.com.

Key words: dermoscopy; artificial intelligence; skin cancer.

Conflict of interest: the authors have no conflict of interest to declare.

Ethics approval and consent to participate: the study was conducted in accordance with the ethical principles of the Declaration of Helsinki. All patients had previously provided written informed consent for the anonymous use of their clinical and dermoscopic images for research purposes.

Availability of data and materials: the datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

Abstract

Artificial intelligence (AI)-based dermoscopic analysis is increasingly being used to assist dermatologists in the early detection and monitoring of pigmented skin lesions. However, discrepancies between AI-generated assessments and clinical evaluation may occur. This retrospective study aimed to evaluate the level of agreement between dermatologist assessment and AI-generated analysis in the dermoscopic evaluation of skin lesions. The study included 71 dermoscopic cases examined at two time points between September 2024 and January 2026 at a dermatologic clinic in Tirana, Albania. Dermoscopic images were acquired using a digital dermoscopy platform equipped with an integrated AI algorithm for automated lesion analysis. Agreement between the dermatologist and the AI system was observed in 58 cases (81.7%), whereas 13 cases (18.3%) demonstrated discordance. The 95% confidence interval for the agreement rate was 72.7%-90.7%, indicating a high level of concordance between clinical assessment and AI-based evaluation. These findings suggest that AI-assisted dermoscopic analysis demonstrates substantial agreement with clinical evaluation and may serve as a valuable supportive tool in dermatologic practice. Nevertheless, the observed discordant cases highlight the continuing importance of clinical expertise and established dermoscopic criteria in guiding patient management. AI systems should therefore be regarded as complementary decision-support tools rather than independent diagnostic systems.

Introduction

Digital dermoscopy has become an indispensable tool in modern dermatology for the identification, documentation, and monitoring of skin lesions. By enabling the acquisition and storage of high-resolution dermoscopic images, it facilitates longitudinal follow-up and enhances the early detection of malignant lesions, particularly melanoma.[1] Early diagnosis remains crucial, as the global incidence of melanoma continues to rise, and patient prognosis is strongly associated with detection at an early stage.[1,2] Despite its well-established diagnostic value, dermoscopic interpretation remains highly dependent on the expertise and experience of the clinician, introducing an inherent degree of subjectivity into the diagnostic process.[2] In recent years, artificial intelligence (AI), particularly convolutional neural network-based algorithms, has demonstrated considerable potential in the automated analysis of dermatologic images, achieving diagnostic accuracies comparable to those of experienced dermatologists in several studies.[2,3] Consequently, AI-based technologies are increasingly being integrated into digital dermoscopy platforms to support lesion assessment, risk stratification, and clinical decision-making.[4] Nevertheless, several challenges continue to limit the implementation of AI in routine clinical practice. Variability in image quality, differences in clinical context, and potential discrepancies between algorithm-generated assessments and dermatologist interpretation may all

influence diagnostic performance.[3,5] Comparative studies evaluating the agreement between clinician assessment and AI-generated analysis are therefore essential to better define the strengths and limitations of these technologies and to clarify their role as decision-support tools in contemporary dermatology.[3,5]

Aims

The aim of this retrospective study was to evaluate the level of agreement between the dermatologist's clinical assessment and the AI-generated analysis provided by the FotoFinder system in the interpretation of dermoscopic images. Additionally, the study sought to identify and analyze potential discrepancies between the two approaches in the longitudinal evaluation of skin lesions.

Materials and Methods

This retrospective study included 71 patients who underwent digital dermoscopic examination of pigmented skin lesions at two time points between September 2024 and January 2026 at a referral dermatology clinic in Tirana, Albania. The study comprised pigmented skin lesions routinely evaluated during digital dermoscopic follow-up, including melanocytic nevi, dysplastic nevi, seborrheic keratoses, pigmented actinic keratoses, and other clinically relevant pigmented lesions. The primary objective of the study was not to establish histopathological diagnoses but to compare the dermatologist's overall clinical assessment with the AI-generated evaluation provided by the FotoFinder system. Data were retrieved from the FotoFinder digital dermoscopy database. The study was conducted in accordance with the ethical principles of the Declaration of Helsinki. All patients had previously provided written informed consent for the anonymous use of their clinical and dermoscopic images for research purposes. Patient confidentiality was strictly maintained throughout the study, and all data were analyzed in an anonymized format. Dermoscopic images were acquired using the FotoFinder digital dermoscopy platform, which enables the acquisition of high-magnification images (20×-40×) and longitudinal comparison of skin lesions. Images were obtained at baseline and follow-up examinations, and the integrated AI algorithm generated a risk score for each time point. Quantitative parameters automatically recorded by the system included lesion area (mm^2), perimeter, two principal diameters, aspect ratio, baseline score, and follow-up score. These data were exported into a structured Microsoft Excel dataset. Additional variables included patient sex, clinical diagnosis, and anatomical location of the lesion. Initially, each case was independently evaluated by an experienced dermatologist based on established dermoscopic criteria, lesion morphology, and clinical follow-up findings. The same lesion was subsequently analyzed using the FotoFinder AI system. Concordance between the two assessments was recorded as a binary variable ("Agreement").

Agreement was defined as consistency between the dermatologist's overall assessment and the AI-generated evaluation with respect to lesion risk and follow-up interpretation. Cases were classified as "Yes" when both assessments were concordant and "No" when they differed. Accordingly, the analysis focused on overall agreement between physician assessment and AI-generated evaluation rather than on concordance across specific diagnostic categories. Patients were eligible for inclusion if dermoscopic images of the same lesion were available at both baseline and follow-up examinations, together with complete clinical, morphometric, and AI-generated data. Cases with missing follow-up images, incomplete datasets, duplicate records, or images of inadequate quality due to poor focus, insufficient illumination, motion artifacts, or incomplete lesion visualization were excluded from the analysis. A representative example of longitudinal follow-up analysis (3 July 2025 and 14 January 2026) is presented in Figure 1. The lesion, located on the anterior aspect of the right thigh, exhibited stable dermoscopic morphology, with no significant changes in surface area or maximum diameter and only minor variations in perimeter and minimum diameter. Images were acquired under identical conditions (20× magnification and standardized illumination). The AI-generated risk score remained low at both examinations, indicating lesion stability.

Initially, each case was evaluated by an experienced dermatologist based on established dermoscopic criteria and clinical judgment. Subsequently, the same lesion was analyzed using the FotoFinder system, which generated a probabilistic risk assessment based on AI algorithms. The comparison between these two evaluations was recorded in the dataset as a separate variable ("Agreement"), indicating whether the dermatologist's assessment and the AI-generated evaluation were concordant (Yes) or discordant (No).

Results

Among the 71 analyzed cases, agreement between the dermatologist's evaluation and the AI-generated assessment was observed in 58 cases, corresponding to an overall agreement rate of 81.7%, whereas 13 cases (18.3%) demonstrated discordance. The 95% confidence interval for the agreement rate was 72.7%-90.7%, indicating a high level of concordance between clinical assessment and AI-based analysis. These findings suggest that, in the majority of cases, the AI system reached conclusions consistent with those of the dermatologist. Nevertheless, the results also demonstrate that AI cannot yet provide a consistently reliable assessment when used independently. Accordingly, it should be regarded primarily as a decision-support tool rather than a standalone diagnostic system. A total of 13 discordant cases were identified between the dermatologist's assessment and the AI-generated evaluation. Several factors may account for these discrepancies. Technical factors, including image quality (focus and illumination), lesion location, and inaccurate identification of lesion borders or

principal axes by the AI system, may influence the generated results. Furthermore, temporal changes occurring during lesion follow-up may affect comparative image interpretation. Another important consideration is that AI algorithms are trained on specific image datasets, which may not fully represent the wide spectrum of skin lesions encountered in routine clinical practice. The following cases illustrate representative examples of discordance between the dermatologist's evaluation and the assessment generated by the FotoFinder AI system.

Case 1: incorrect lesion delineation

The lesion was located on the frontal scalp (head front), an anatomical site characterized by substantial chronic sun exposure. Dermoscopic images were acquired at 20× magnification during two separate examinations. Based on clinical and dermoscopic evaluation, the lesion appeared as a small, brown pigmented lesion with relatively well-defined borders. Although the AI system generated a high-risk score at both time points, suggesting a potential risk of malignancy, the dermatologist elected to continue clinical follow-up rather than proceed with surgical excision, based on established dermoscopic criteria and overall clinical assessment. As shown in Figure 2, the contour delineated by the AI system (dark outline) encompasses an irregular area of the surrounding skin, whereas the actual pigmented lesion lies partially outside the selected region. Consequently, the elevated AI-generated risk score does not reflect the true morphological characteristics of the lesion but rather appears to result from an error during lesion detection and segmentation. This segmentation error may have been influenced by the low contrast between the lesion and the surrounding skin, as well as by subtle variations in skin pigmentation, both of which may interfere with the algorithm's ability to accurately identify lesion boundaries.

Case 2: decreased AI risk score due to a segmentation error

This case involved a male patient diagnosed with a dysplastic nevus, with the lesion located on the posterior aspect of the right lower leg. The baseline examination was performed on 05 May 2025, followed by a follow-up examination on 16 September 2025 (Figure 3). Morphometric analysis demonstrated an approximately 20% increase in lesion surface area and dimensions. However, visual comparison of the dermoscopic images suggested that this apparent increase did not represent true lesion enlargement. In the baseline image, the lesion margins were not completely captured, resulting in partial exclusion of the lesion during the segmentation process. Consequently, the initial lesion area was artificially underestimated, as only the region with the most intense pigmentation was segmented, whereas the peripheral portion of the lesion was omitted. This explains why the baseline image, in which pigmentation appeared more pronounced because of the segmentation error, received a high AI

risk score, whereas the follow-up image did not, leading to discordance between the dermatologist's assessment and the AI-generated evaluation. Although the AI system assigned a high-risk score to the lesion at baseline, suggesting a possible diagnosis of melanoma, the dermatologist, based on established dermoscopic criteria and clinical judgment, elected to monitor the lesion rather than perform surgical excision. Follow-up evaluation confirmed the appropriateness of this management strategy, demonstrating that the initial AI assessment had been influenced by a technical inconsistency during image acquisition and lesion segmentation.

Case 3: interpretative discordance

This case involved a female patient, born in 1964, diagnosed with a dysplastic nevus, with the lesion located on the anterior upper trunk. The baseline examination was performed on 9 August 2025, and the follow-up examination on 17 October 2025 (Figure 4). Both dermoscopic images were acquired at 20× magnification. Quantitative analysis demonstrated an approximately 20% increase in lesion surface area and diameters, accompanied by an increase in the AI-generated risk score from 84% to 93%. However, visual comparison of the dermoscopic images revealed that the lesion remained essentially unchanged at both time points, with very similar pigmentation, an unchanged internal dermoscopic structure, and well-preserved lesion borders.

Discussion

AI-assisted dermoscopy has become increasingly integrated into dermatologic practice as a valuable tool for supporting the early detection of skin cancer. In the present study, agreement between the dermatologist's evaluation and the AI-generated assessment was observed in 81.7% of cases, indicating a high level of concordance between clinical interpretation and algorithm-based analysis. This finding is consistent with previous studies demonstrating that contemporary deep-learning algorithms can achieve diagnostic performance comparable to that of experienced dermatologists in the classification of pigmented skin lesions. A landmark study by Esteva *et al.*[3] demonstrated that a deep convolutional neural network was capable of classifying skin cancer with a diagnostic performance comparable to that of dermatologists when trained on a large dataset of clinical and dermoscopic images. Similarly, Haenssle *et al.*[6] reported that AI-based algorithms achieved diagnostic accuracy comparable to that of experienced dermatologists for the detection of melanoma using dermoscopic images. Collectively, these findings support the expanding role of AI systems as valuable adjuncts in dermatologic diagnosis. Despite the high level of concordance observed in the present study, discordance between the dermatologist and the AI system was identified in 18.3% of cases, highlighting an important limitation of AI-assisted diagnostic tools. Several factors may account for

these discrepancies. Technical issues related to image acquisition, including suboptimal focus, illumination artifacts, and variations in magnification, may adversely affect image quality and consequently influence algorithm performance. Furthermore, lesions located in anatomically challenging regions or incompletely captured during image acquisition may be associated with inaccurate lesion segmentation and border detection by the AI system. The discordant cases analyzed in this study provide important insights into the practical limitations of AI-assisted dermoscopy. In several cases, the discrepancy was primarily attributable to segmentation errors, whereby inaccurate lesion delineation resulted in misleading AI-generated risk scores. In other cases, discordance appeared to arise from technical variations during image acquisition or from the algorithm assigning disproportionate importance to minor quantitative changes that were not considered clinically meaningful by the dermatologist. These observations demonstrate that AI performance may be influenced not only by lesion characteristics but also by technical and procedural factors encountered in routine clinical practice. Another important consideration is the composition of the datasets used to train AI algorithms. Machine-learning models rely heavily on large, annotated image datasets, which may not fully capture the diversity of skin lesions encountered in everyday clinical practice. Variations in skin phototype, lesion morphology, and image acquisition conditions may therefore limit the generalizability of AI algorithms. As noted by Celebi *et al.*,[7] algorithm performance may decline when applied to datasets that differ from those used during model development, underscoring the need for continuous validation in real-world clinical settings. Beyond the technical explanations for the observed discordance, the potential clinical implications deserve careful consideration. In the present study, none of the discordant cases resulted in a missed melanoma or other malignant lesion during the available follow-up period. Several lesions classified as high risk by the AI system were considered clinically benign or stable based on dermoscopic evaluation and longitudinal clinical assessment by the dermatologist. Had management decisions relied solely on AI-generated risk scores, some patients might have undergone unnecessary surgical excision, additional diagnostic investigations, or more intensive follow-up than clinically warranted. Such outcomes could increase healthcare costs while unnecessarily contributing to patient anxiety. Although no clinically significant false-negative cases were identified in this cohort, the possibility of underestimating malignant transformation remains an important concern when AI systems are used without appropriate clinical oversight. These findings emphasize the importance of interpreting AI-generated assessments within the broader clinical and dermoscopic context. The discordant cases presented in this study demonstrate that dermatologist expertise remains indispensable for distinguishing true biological changes from technical artifacts, segmentation errors, and algorithmic overestimation of risk, thereby ensuring appropriate patient management. Clinical context remains a fundamental component of dermatologic decision-making.

Dermatologists integrate dermoscopic findings with patient history, lesion evolution, anatomical location, and other clinical variables that cannot be fully captured by image-based algorithms alone. This comprehensive clinical perspective enables more accurate interpretation of dermoscopic findings and helps avoid unnecessary surgical procedures when lesions remain clinically and dermoscopically stable during follow-up. In several discordant cases included in the present study, longitudinal follow-up confirmed lesion stability or identified technical inconsistencies in the initial image acquisition, thereby supporting the dermatologist's clinical judgment. These observations reinforce the concept that AI should complement, rather than replace, clinical expertise.[8]

Several limitations of this study should be acknowledged. First, the sample size was relatively small, comprising only 71 lesions, which may limit the statistical precision of the findings and reduce the ability to detect less common patterns of disagreement between dermatologist and AI assessments. Second, this was a single-center study conducted at a referral dermatology clinic in Tirana, Albania, which may limit the generalizability of the findings to other populations, healthcare settings, and imaging environments. Third, the retrospective design is inherently susceptible to selection bias and relies on the quality and completeness of previously recorded clinical and dermoscopic data. Finally, histopathological confirmation was not available for all lesions because the primary objective of the study was to evaluate agreement between dermatologist assessment and AI-generated analysis during routine clinical follow-up rather than to determine diagnostic accuracy.

Overall, the findings of this study add to the growing body of evidence supporting the role of AI-based dermoscopic analysis as a valuable adjunct in dermatology, particularly for lesion documentation, longitudinal monitoring, and risk stratification. Nevertheless, the discordant cases identified in this study underscore the current limitations of AI systems and reinforce the indispensable role of dermatologist expertise in the interpretation of dermoscopic findings and the guidance of appropriate clinical management. Future multicenter studies involving larger patient cohorts and longer follow-up periods are warranted to further define the role of AI in routine dermoscopic practice and to improve the reliability and clinical applicability of AI-assisted diagnostic systems.

Conclusions

Overall, the findings of this study further support the growing evidence that AI-based dermoscopic analysis can serve as a valuable supportive tool in dermatology, particularly for lesion documentation, longitudinal follow-up, and risk stratification. Nevertheless, accurate diagnosis and appropriate patient management continue to rely on the clinical expertise of experienced dermatologists. AI should therefore be regarded as a complementary decision-support tool rather than a replacement for clinical judgment.

References

1. Marchetti MA, Cowen EA, Kurtansky NR, et al. Prospective validation of dermoscopy-based open-source artificial intelligence for melanoma diagnosis (PROVE-AI study). NPJ Digit Med 2023;6:127.
2. Kittler H, Pehamberger H, Wolff K, Binder M. Diagnostic accuracy of dermoscopy. Lancet Oncol 2002;3:159-65.
3. Esteva A, Kuprel B, Novoa RA, et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature 2017;542:115-8. Erratum in: Nature 2017;546:686.
4. Tschandl P, Rosendahl C, Akay BN, et al. Expert-Level Diagnosis of Nonpigmented Skin Cancer by Combined Convolutional Neural Networks. JAMA Dermatol 2019;155:58-65.
5. Tschandl P, Rosendahl C, Kittler H. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. Sci Data 2018;5:180161.
6. Haenssle HA, Fink C, Schneiderbauer R, et al. Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists. Ann Oncol 2018;29:1836-42.
7. Celebi ME, Codella N, Halpern A. Dermoscopy Image Analysis: Overview and Future Directions. IEEE J Biomed Health Inform 2019;23:474-8.
8. Daneshjou R, Kovarik C, Ko JM. Towards realization of augmented intelligence in dermatology: advances and future directions. NPJ Digit Med 2021.

Figure 1. Example of a comparison report generated with the FotoFinder system.

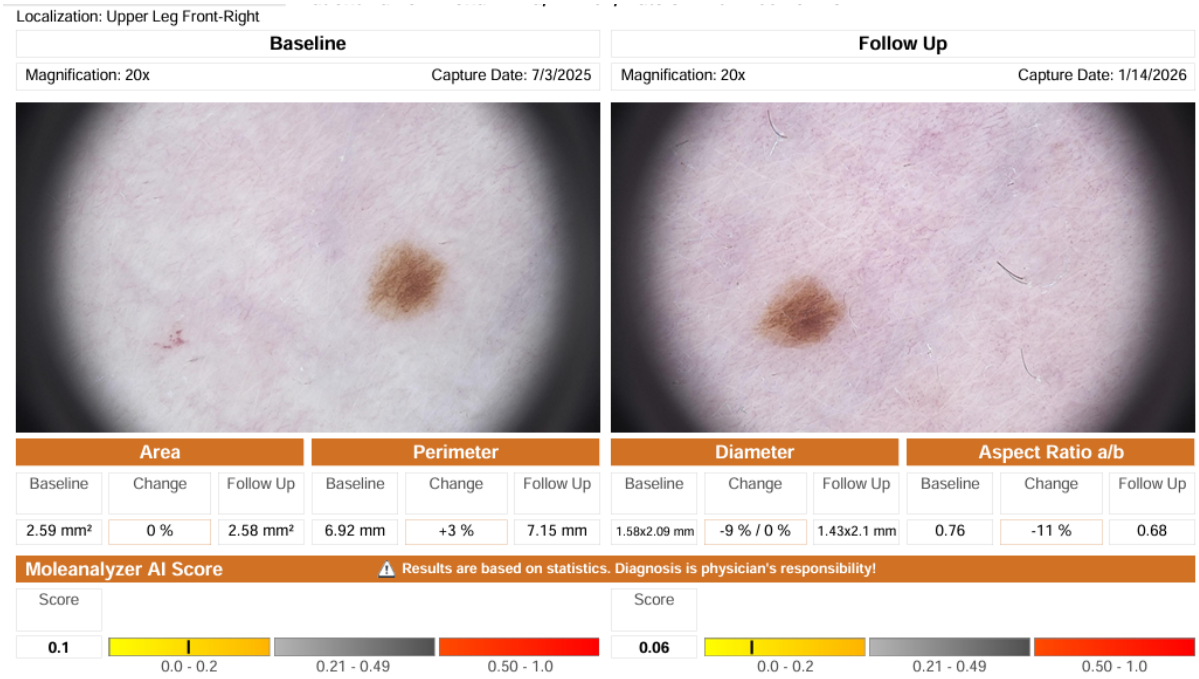

Figure 2. Case 1 showing discordance between the dermatologist's evaluation and the AI-generated assessment.

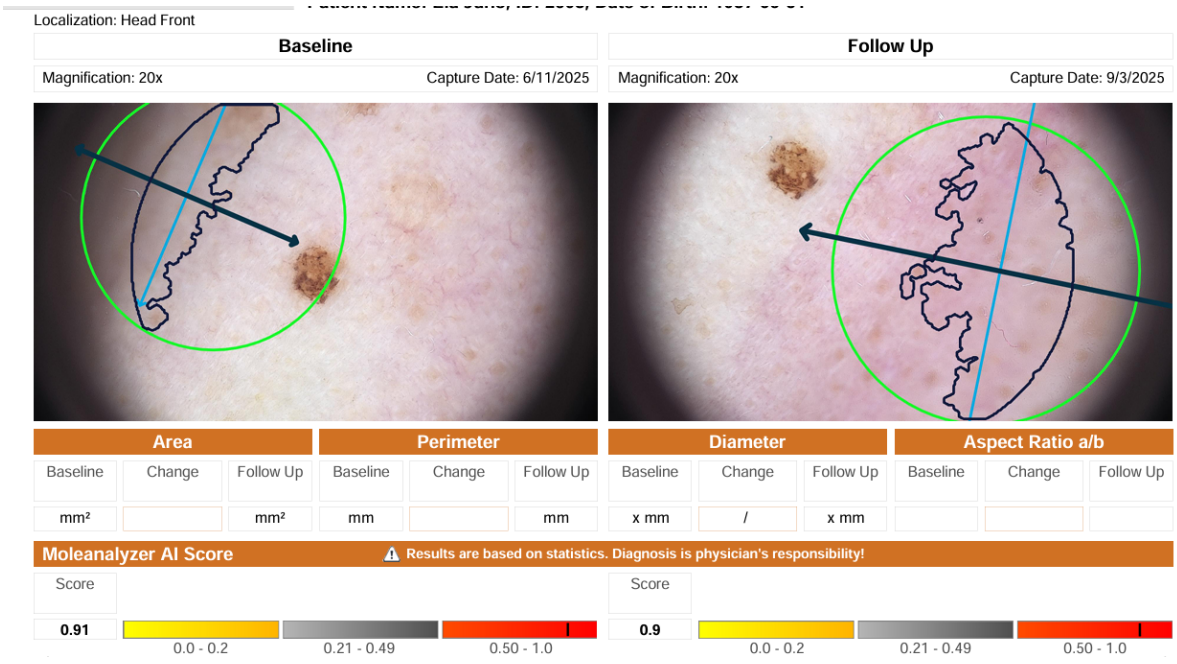

Figure 3. Case 2 showing discordance between the dermatologist's evaluation and the AI-generated assessment.

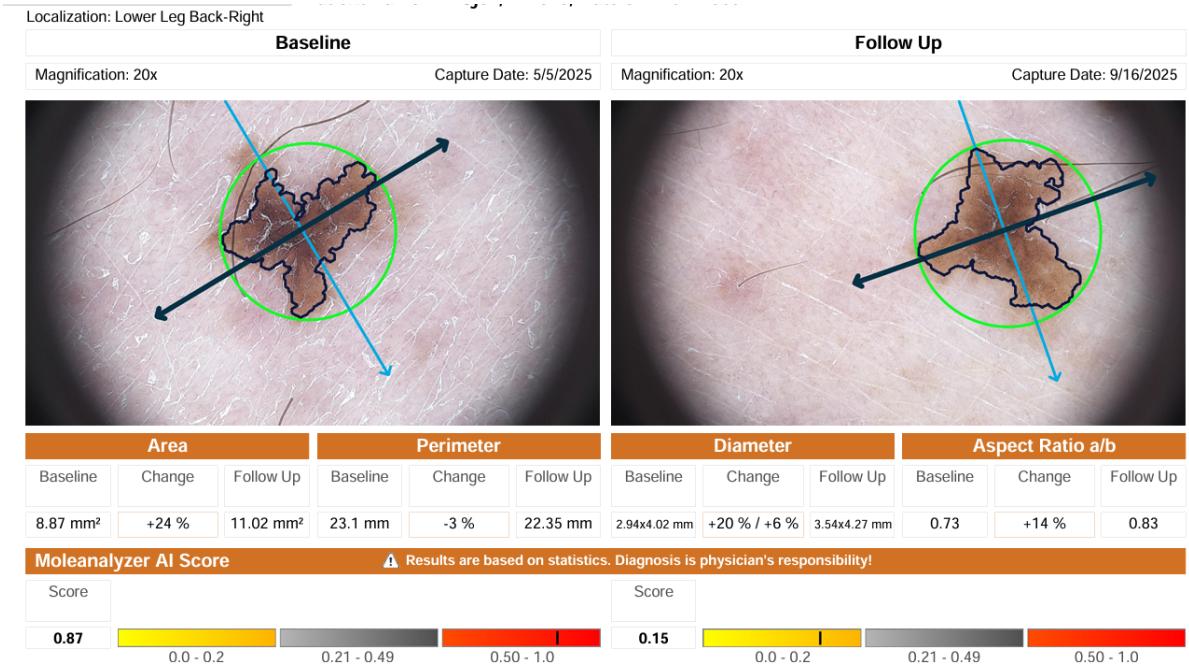

Figure 4. Case 3 showing discordance between the dermatologist's evaluation and the AI-generated assessment.