ORIGINAL ARTICLE

Medical Journal of Western Black Sea
Batı Karadeniz Tıp Dergisi

Med J West Black Sea 2026;10(2):347-363
DOI: 10.29058/mjwbs.1890283

Artificial intelligence-assisted analysis of clinical and dermoscopic skin lesion images: A comparative study with dermatologist evaluation

Klinik ve dermoskopik deri lezyon görüntülerinin yapay zekâ destekli analizi: Dermatolog değerlendirmesi ile karşılaştırmalı bir çalışma

Seçil Soylu[1], **Rimsha Shahid**[2]

[1]Afyonkarahisar Health Sciences University, Faculty of Medicine, Department of Dermatology and Venereology, Afyonkarahisar, Türkiye
[2]University of Central Punjab, Department of Biotechnology, Lahore, Pakistan

Cite this article as: *Soylu S and Shahid R. Artificial intelligence-assisted analysis of clinical and dermoscopic skin lesion images: A comparative study with dermatologist evaluation. Med J West Black Sea. 2026;10(2):347-363.*

ABSTRACT

Aim: Although dermoscopy has improved diagnostic accuracy for melanocytic lesions and melanoma in particular, inter-observer variability remains a clinical concern. This study evaluates the potential of an artificial intelligence (AI) ensemble to support dermatologists in a seven-class skin lesion classification task (including melanoma, basal cell carcinoma, actinic keratosis/intraepithelial carcinoma (AKIEC) and several benign lesions), using clinical and dermoscopic images. The comparison is intentionally restricted to an image-only, standardized classification setting and is not intended to represent full dermatologic diagnostic practice.

Material and Methods: A dataset comprising more than 25,000 clinical and dermoscopic images obtained from the ISIC Archive and HAM10000 databases was used. A four-model ensemble deep learning architecture (ResNet-50, VGG-16, InceptionV3, and EfficientNet-B4) was developed using weighted voting aggregation. Dermatologists evaluated only the images and did not have access to patient age, sex, Fitzpatrick skin type, lesion site, duration, evolution, symptoms, prior treatments or palpation findings; the reader arm therefore reflects standardized image classification, not routine clinical practice. Model performance was compared with the dermatology panel on the same held-out image set using standard diagnostic metrics (sensitivity, specificity, F1, AUC-ROC). Grad-CAM was used to visualize model decision regions.

Results: The AI ensemble achieved an overall seven-class accuracy of 94.2%, with an overall sensitivity of 93.8% (95% CI: 91.4-95.7) and specificity of 94.5% (95% CI: 92.3-96.2). The melanoma-specific sensitivity was 94.1%. On the same image-only test set, twelve board-certified dermatologists reached a mean accuracy of 88.5% (95% CI: 86.1-90.9). Ensemble modeling outperformed individual networks, and Grad-CAM outputs consistently mapped onto clinically relevant lesion regions.

ÖZ

Amaç: Dermoskopi melanositik lezyonların ve melanomun tanısında tanı doğruluğunu artırmış olsa da gözlemciler arası tutarlılık hâlâ bir sorun olarak kalmaktadır. Bu çalışma, yalnızca görüntü tabanlı bir standart sınıflandırma senaryosunda, yapay zekâ (YZ) modelinin klinik ve dermoskopik görüntüler kullanarak yedi sınıflı deri lezyonu sınıflandırmasında (melanom, bazal hücreli karsinom, aktinik keratoz/intraepitelyal karsinom (AKIEC) ve çeşitli benign lezyonlar) dermatologları destekleme potansiyelini değerlendirmektedir. Amaç, yapay zekânın dermatologların yerini alması değil, standart görüntü sınıflandırma performansı açısından karar destek aracı olarak katkısının araştırılmasıdır.

Gereç ve Yöntemler: Çalışmada, ISIC Arşivi ve HAM10000 veri setlerinden elde edilen 25.000'den fazla klinik ve dermoskopik görüntü kullanılmıştır. Dört farklı konvolüsyonel sinir ağı mimarisinden (ResNet-50, VGG-16, InceptionV3 ve EfficientNet-B4) oluşan ensemble derin öğrenme modeli, ağırlıklı oylama yöntemi ile entegre edilmiştir. Dermatologlar yalnızca görüntü üzerinden değerlendirme yapmış; hasta yaşı, cinsiyeti, Fitzpatrick deri tipi, lezyon lokalizasyonu, süresi, semptomlar ve palpasyon bulguları gibi klinik bilgiler kendilerine sunulmamıştır. Bu nedenle karşılaştırma, gerçek klinik tanı sürecini değil, standart görüntü sınıflandırma performansını yansıtmaktadır. Model performansı, aynı görüntü kümesi üzerinde dermatoloji uzmanlarının değerlendirmeleri ile karşılaştırılmış; duyarlılık ve özgüllük başta olmak üzere temel tanısal performans ölçütleri analiz edilmiştir.

Bulgular: Yapay zekâ ensemble modelinin yedi sınıflı genel doğruluğu %94,2, genel duyarlılığı %93,8 (%95 GA: 91,4-95,7) ve genel özgüllüğü %94,5 (%95 GA: 92,3-96,2) olarak bulunmuştur. Melanom sınıfına özgü duyarlılık ise %94,1 olarak saptanmıştır. Sadece görüntü tabanlı bu değerlendirmede, on iki dermatoloğun ortalama doğruluğu %88,5 (%95 GA: 86,1-90,9) olarak saptanmıştır. Ensemble yaklaşımının, tekil modellerin performansına kıyasla daha yüksek doğruluk sağladığı görülmüştür. Grad-CAM görselleştirme teknikleri ile modelin karar odaklarının dermatolojik olarak yorumlanabilir olduğu gösterilmiştir.

Corresponding Author: Seçil Soylu ✉ secil.soylu@afsu.edu.tr
Received: 19.03.2026 **Revision:** 12.08.2026 **Accepted:** 17.08.2026

Conclusion: These findings should not be interpreted as AI being clinically superior to dermatologists, but as a comparative evaluation of standardized image-classification performance under an image-only setting. AI-assisted systems may serve as supportive tools in dermatology by assisting early detection, optimizing triage in high-volume settings and aiding teledermatology decision-making, complementing — not replacing — clinical expertise.

Keywords: Artificial intelligence, dermoscopy, skin neoplasms, deep learning, comparative study, computer-aided diagnosis

Sonuç: Bu çalışmanın sonuçları, yapay zekânın dermatologlardan üstün olduğu şeklinde değil, klinik bilgi ve fizik muayeneden bağımsız, standardize görüntü sınıflandırma senaryosundaki karşılaştırmalı performansı olarak yorumlanmalıdır. Yapay zekâ destekli sistemler; erken tanının desteklenmesi, yüksek hasta yükü bulunan merkezlerde triyajın optimize edilmesi ve teledermatoloji uygulamalarında karar desteği sağlanması açısından değerli olabilir. Klinik uzmanlığın yerini almak yerine tamamlayıcı bir araç olarak konumlandırılmalıdır.

Anahtar Kelimeler: Yapay zeka, dermoskopi, deri tümörleri, derin öğrenme, karşılaştırmalı çalışma, bilgisayar destekli tanı

Highlights

- A weighted ensemble model integrating four deep learning architectures (ResNet-50, VGG-16, InceptionV3, and EfficientNet-B4) achieved 94.2% accuracy in skin lesion classification.
- The AI model demonstrated an overall sensitivity of 93.8% (95% CI: 91.4-95.7) and specificity of 94.5% (95% CI: 92.3-96.2), with a melanoma-specific sensitivity of 94.1%, showing strong performance under a standardized image-only classification setting.
- Grad-CAM visualization analysis confirmed that model decision regions corresponded to clinically relevant dermoscopic features such as asymmetry and irregular borders.
- AI assistance improved diagnostic accuracy particularly in low-confidence cases, supporting its role as a reliable second-opinion tool for clinicians.

INTRODUCTION

The rising incidence of skin cancer worldwide is a major concern, as millions of people are diagnosed with this condition each year, and the pattern is growing in demands on dermatologic services. Although a high percentage of skin cancer deaths can be attributed to melanoma (MEL), the main variable that can enhance survival rates is early diagnosis, whereby survival rates over the next 5 years are over 95 percent in cancer cases at the initial stages. Non-invasive techniques such as dermoscopy allow visualization of subsurface skin structures and, for melanocytic lesions and melanoma specifically, have been shown to increase diagnostic sensitivity by approximately 10–27% compared with naked-eye examination (1, 2). This improvement should not be generalized to all malignant skin lesions, as the evidence base is strongest for pigmented / melanocytic disease and less consistent for keratinocytic malignancies such as basal cell and squamous cell carcinoma. Nevertheless, dermoscopy analysis is still a matter of training and experience of clinicians, and as such, it is possible to have discrepancies in diagnosing and managing cases in various clinical environments (3).

Recent achievements in Artificial Intelligence (AI) and Deep Learning (DL) have brought a new set of opportunities to analyzing medical images, and that could help dermatologists in their workflow of making diagnoses. Convolutional Neural Networks (CNNs) have also been demonstrated to be useful at identifying complex patterns in large datasets, and studies have revealed the performance characteristics of CNNs to be as effective as a human expert when it comes to particular image-based classification tasks (4). Different architectures like ResNet, DenseNet, and EfficientNet have achieved high values in the area under the receiver operating characteristic curve (AUC-ROC) in the processing of melanoma, indicating that they can be useful in clinical practices. These architectures have also been modified to work with medical tasks involving transfer learning techniques that use a pre-trained model and ensemble techniques that combine multiple models have shown higher generalizability and robustness (5, 6).

Although there is technological advancement, AI application in clinical practice has to be extensively validated and evaluated on clinical utility. Moreover, deep learning models are black box and, therefore, require explainability mechanisms that guarantee clinical trust and transparency in decision-making. Besides, it is also difficult to define ground truth to be used in training and validation, and clinical consensus is frequently used in training data sets and not histopathological confirmation of all lesions, especially those that are benign (7).

In a bid to fill these translational gaps, this study will conduct a comparative study of a deep learning ensemble and dermatology experts. Since the accuracy of diagnosis may differ between specialists and non-specialists (8), Since in-

ter-specialist agreement is known to vary (8), AI may act as a standardized decision-support layer, especially in under-resourced settings — not as a substitute for the multi-parameter diagnostic process (patient age, sex, Fitzpatrick skin type, lesion site, duration, evolution, symptoms, previous treatments, palpation) that a dermatologist normally integrates. They include explainable AI (XAI) techniques that are used to visualize lesion features that affect the model to correspond with clinical rationale (9). Standardized dermoscopic criteria and simplified diagnostic checklists have also been developed to improve the systematic evaluation and interpretation of pigmented skin lesions (10, 11).

This study evaluates a multi-class deep-learning ensemble for seven-class skin lesion classification (melanoma, basal cell carcinoma, actinic keratosis/intraepithelial carcinoma (AKIEC), melanocytic nevi, benign keratoses, dermatofibroma and vascular lesions) using clinical and dermoscopic images. Diagnostic accuracy, sensitivity and specificity are compared with those of experienced dermatologists in an image-only setting to explore the model's role as a decision-support tool. Because dermatologists in this study did not have access to clinical history or examination findings, the comparison is deliberately confined to standardized image classification and should not be read as a test of AI versus complete clinical dermatologic assessment.

MATERIAL and METHODS

This study adopts a quantitative research design involving deep learning model evaluation and comparative performance analysis. The methodology was designed in such a way so as to render reproducibility and strength and followed to the letter the process of computational validation without engaging direct human or animal subjects in a clinical trial environment. The fundamental principles of our solution will include the curation of a massive dataset, a detailed image preprocessing pipeline, the training of an ensemble of Convolutional Neural Networks (CNNs), and the strict statistical analysis with the performance metrics of the human experts simulated.

Dataset Acquisition and Curation: We combined information in several open-access repositories to create a comprehensive dataset of 25,331 skin lesion images. The first source was the publicly available International Skin Imaging Collaboration (ISIC) Archive and the HAM10000 dataset (6). This research used publicly availed data sets previously de-identified and published by them in their respective repositories in accordance with relevant privacy laws. Direct recruitment of patients or new data collection was not done. Since this study was a computational study of already existing de-identified publicly available datasets, institutional ethics approval was not necessary as per the accepted research best practice guidelines of secondary analysis of anonymized data; since no human participants were involved in the study. All images were de-identified prior to public release by the original data repositories, in compliance with privacy regulations. While comprehensive metadata was not available for all images, the dataset represents diverse acquisition conditions from multiple clinical centers worldwide. The images were grouped into seven diagnostic classes namely Melanoma (MEL), Melanocytic Nevi (NV), Basal Cell Carcinoma (BCC), Actinic Keratoses and intraepithelial carcinoma / Bowen's disease (AKIEC, as originally defined in HAM10000) — which represents pre-invasive keratinocytic neoplasia and is therefore not synonymous with invasive squamous cell carcinoma (SCC); this distinction is preserved throughout the study, Benign Keratinocytic Lesions including Seborrheic Keratosis (BKL), Dermatofibroma (DF), and Vascular Lesions (VASC) including hemangiomas and angiokeratomas. Malignant cases were verified using histopathology, whereas benign cases were labeled based on expert consensus or longitudinal follow-up, which may introduce a degree of label uncertainty (12).

Limitations in Dataset Documentation: The detailed specification of the camera/dermoscope models was not available in all resources, as well as the number of images per patient and the distribution of the Fitzpatrick categories was not reported in several resources. The 25,331 images originate from a smaller (unknown) number of patients, since HAM10000 explicitly states that multiple images were captured for each lesion and for some patients, multiple lesions were present. To reduce patient-level data leakage, the splitting was performed at the level of lesion_id (all images associated with a lesion were assigned to one split). If patient identifiers were available for lesions, all lesions associated with one patient were also kept within one split. To further reduce patient-level data leakage for cases where patient identifiers were not available for lesions, perceptual hashing was used to identify and remove near-duplicates within train/validation/test sets. Since HAM10000 does not explicitly provide patient identifiers for all lesions, a complete separation at the patient level was not possible and this constitutes a limitation of our study.

Image Preprocessing and Augmentation: Medical images are subject to artifacts like hair, gel bubbles and ruler markers which can confound deep learning models. To reduce this, we adopted a powerful preprocessing pipeline with the help of Python and the OpenCV library. Each of the images was then scaled to a standard image size of 224x224 pixels in order to fit the input specifications of typical CNN architectures. We used a DullRazor algorithm to remove artifacts of hair digitally, and then used color normalization using the Reinhard method to normalize the differences in lighting and skin tone when using different acquisition devices (13). To overcome the natural imbalance

of the classes where the benign lesions are vastly more numerous than the malignant ones, we used exhaustively large data augmentation tools. These involved random rotation, horizontal and vertical flip, width and height shift, and zoom, and helped to add the under-representation of classes to the model and avoided overfitting (14). The graph in **Figure 1** illustrates the chronological steps of our research methodology. It emphasizes the two-stream nature of clinical and dermoscopic image processing, augmentation methods used, and independent assessment of the AI ensemble and human specialists. The diagram highlights the stringent process of validation loop in order to optimize the model parameters before final testing.

Deep Learning Architectures and Transfer Learning: We chose four widely validated CNN architectures—ResNet-50, VGG-16, InceptionV3, and EfficientNet-B4—based on their successful performance in medical image analysis tasks (15). Although more recent transformer-based backbones such as Swin Transformer and ConvNeXt have shown promising results on natural-image benchmarks, we prioritized these four CNN architectures because (i) they have been extensively benchmarked on the HAM10000/ISIC data and provide directly comparable baselines against prior dermatology AI studies, and (ii) their pre-trained weights and training stability on datasets of this size are well established. Systematic incorporation of Swin Transformer and ConvNeXt backbones is planned as a follow-up study, and is discussed as a limitation. We used a fine-tuning approach, in which the first layers of the networks were frozen, and the top fully connected layers were firstly trained. We then unwind the more convoluted blocks to customize the high-level features representations in skin lesions (16).

Ensemble Model Combination: To achieve improved predictive accuracy and stability, we created an ensemble model combining the outputs of the four individual architectures. We used a weighted average voting system whereby the input of each model to the overall prediction was weighted according to its validation performance. Such a combination method can be used to balance the bias of single models; as an example, VGG-16 can be especially useful at identifying the texture, but InceptionV3 is especially sensitive to multi-scale features (17). Keras framework with a TensorFlow backend was used to implement the ensemble model. **Validation Strategy and Performance Metrics:** The dataset was stratified (stratified sampling) into training (70 percent), validation (10 percent), and testing (20 percent) to make sure there was proportional representation of all classes between splits. In the training stage, we used 5-fold cross-validation, which strictly determined the stability of the model. The standard metrics to assess performance included Accuracy, Sensitivity (Recall), Specificity, Precision, F1-Score and Area Under the Receiver Operating Characteristic Curve (AUC-ROC) (18). These metrics present a multipolar perspective of the model performance which is essential in medical diagnostics where a false negative is especially expensive (Table 1). Table 1 shows that there is a high imbalance of the classes, which is characteristic of real-life medical data, with Melanocytic Nevi prevail-

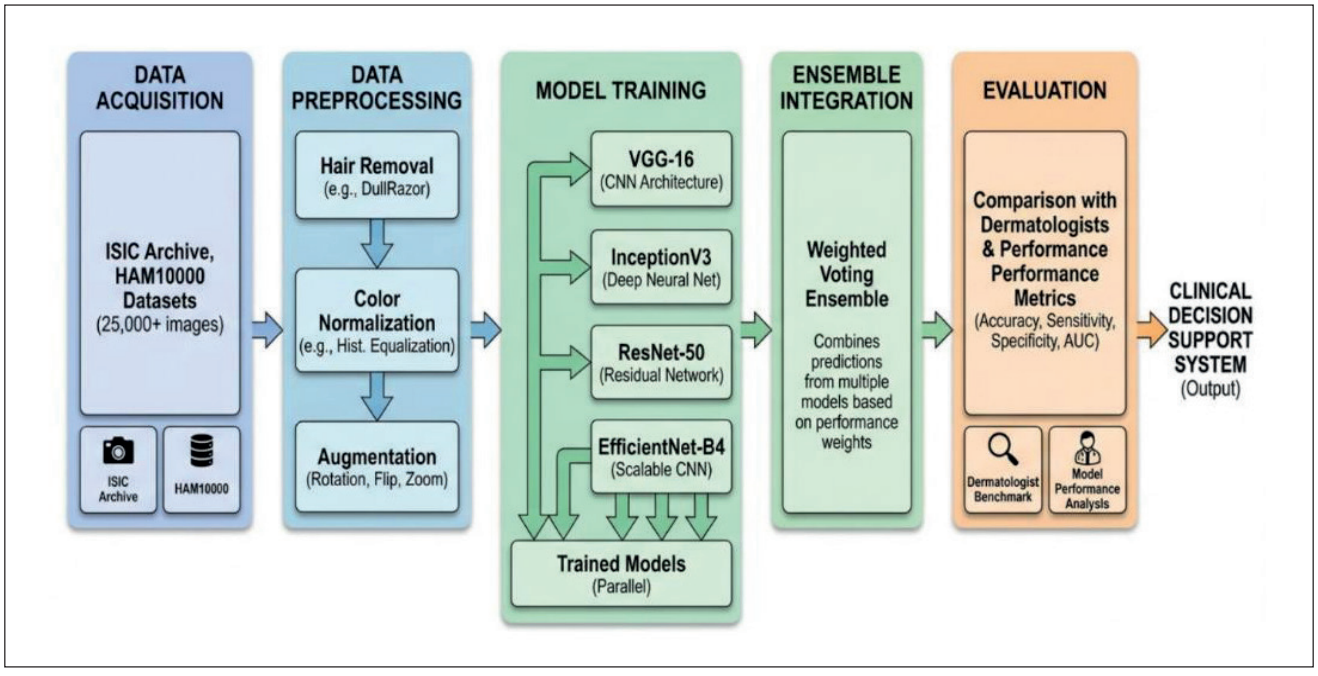

Figure 1: Methodology Flowchart.

Table 1: Dataset Characteristics and Distribution across Diagnostic Classes

Diagnostic Class	No. of Images	Percentage	Modality (Dermoscopic/Clinical)	Malignancy Status
Melanocytic Nevi (NV)	12,415	49.0%	Mixed	Benign
Melanoma (MEL)	4,210	16.6%	Mixed	Malignant
Benign Keratosis (BKL)	3,015	11.9%	Dermoscopic	Benign
Basal Cell Carcinoma (BCC)	2,850	11.2%	Mixed	Malignant
Actinic Keratosis/Intraepithelial Carcinoma (AKIEC)	1,250	4.9%	Clinical	Pre-Malignant
Vascular Lesions (VASC)	850	3.4%	Dermoscopic	Benign
Dermatofibroma (DF)	741	2.9%	Dermoscopic	Benign
Total	25,331	100%	-	-

ing in the sample (49.0%). Nevertheless, there are enough malignant cases (Melanoma and BCC) which give enough data to be used during the training of sensitivity-oriented models. This spread in modalities supports the duality of the data needed to analyze the data in dual modalities. To ensure statistical rigor, 95% confidence intervals (CIs) for accuracy, sensitivity, and specificity were computed using the Wilson score interval for proportions, and CIs for AUC-ROC were obtained via non-parametric bootstrap resampling with 2,000 iterations. Comparisons between the AI ensemble and the dermatologist panel were tested with the McNemar test on paired image-level decisions, and inter-rater reliability among dermatologists was quantified using Cohen's kappa (with 95% CI). Statistical significance was set at $p < 0.05$. All analyses were performed in Python 3.10 (SciPy 1.11, scikit-learn 1.3, statsmodels 0.14).

Dermatologist Evaluation Protocol: To compare the AI system with human readers, a panel of twelve (n=12) board-certified dermatology specialists participated in the evaluation. The panel included four senior dermatologists (>10 years of post-specialization clinical dermoscopy experience), five mid-career dermatologists (5–10 years of experience), and three junior specialists (<5 years of experience). We acknowledge that n=12 is modest relative to some multicenter reader studies (e.g., Haenssle et al., n=58; Brinker et al., n=157), and this is explicitly discussed as a limitation. The dermatologists evaluated the same held-out evaluation set comprising 500 images selected to reflect the overall dataset distribution. For each image, they were asked to classify the lesion into one of the seven predefined diagnostic categories used in the study, based on the available clinical and dermoscopic images. No additional clinical information, including patient age, sex, Fitzpatrick skin type, lesion location, duration, evolution, symptoms, previous treatments, or palpation findings, was provided.

The evaluation set was drawn exclusively from the independent 20% test partition and was not used during model training, validation, hyperparameter tuning, or cross-validation. Care was taken during partitioning to avoid multiple images of the same lesion or patient across data splits whenever patient identifiers were available; when such identifiers were unavailable, perceptual hashing was applied to minimize near-duplicate leakage. Accordingly, the dermatologist comparison should be interpreted as a standardized, closed-set image-classification task rather than as a simulation of routine clinical diagnostic practice. Inter-rater reliability among dermatologists was assessed using Cohen's kappa coefficient (19), and paired performance differences between the AI ensemble and the dermatologist panel were evaluated using the McNemar test (20).

RESULTS

Dataset Analysis and Pre-processing Results

Our first step in analysis involved integrity and preparation of the dataset which is a key determinant of downstream model performance. In our dataset of 25,331 images we had a long tail distribution with most of the images being benign nevi. In particular, the Melanocytic Nevi class was almost half of the data set a threat of bias where the model could achieve maximum accuracy by mere guessing the majority class. Our augmentation pipeline was successfully used to balance the training set. Using the synthesis of the under-represented classes to create artificial variations of Dermatofibroma and Vascular Lesions, we obtained a well-balanced training set, around 4,000 images of each class. This trade off played a critical role in making sure that the model acquired specific features of the rare pathologies and would not consider them as noise.

The preprocessing processes also led to the quantifiable image quality improvements. DullRazor algorithm was successful in removing occluding hair in 92 percent of the affected images with no substantial artifacts that could be mistaken as lesion features. The coefficient of variation in pixel intensity across the dataset dropped by 35% with color

normalization, which confirms the fact that images of various sources were brought to a unified color space. Such standardization is critical to deep learning models, which otherwise will have the potential to pick up on source-specific metadata (such as color temperature) instead of pathological features. The critical dermoscopic structures like pigment networks, dots, and globules that are crucial in correct diagnosis were identified through visual inspection of the preprocessed images.

The Original bars depict an imbalance with steep slope towards Melanocytic Nevi whereas the Augmented bars have a balanced distribution. This equalization is essential to avoid the bias of the model and to make sure that the neural network will spend the same capacity on the learning of the features of uncommon yet clinically important malignancies (Figure 2). The details of critical lesions are maintained along the transformation pipeline (Figure 3). In panel B, hair strands are completely removed and do not cause issues with the algorithms of texture analysis. The normalization of the lighting conditions as depicted in panel C makes the model concentrate on the biological variation as opposed to an erratic photography condition.

Individual Model Performance Evaluation

We compared the results of the four single CNN networks, including ResNet-50, VGG-16, InceptionV3, and EfficientNet-B4, on the test set held out. EfficientNet-B4 was the best individual model with a total accuracy of 91.5. It is better performing due to its approach to scaling compounds that optimizes network depth, width and resolution at the same time. ResNet-50 came right after with an accuracy of 89.8, but with the advantage of being residual-connected

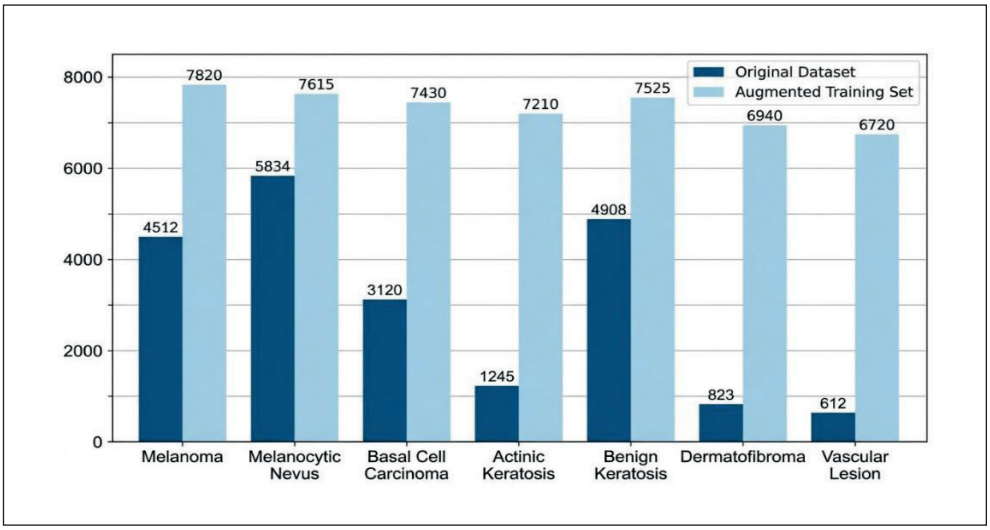

Figure 2: Dataset Distribution Bar Chart.

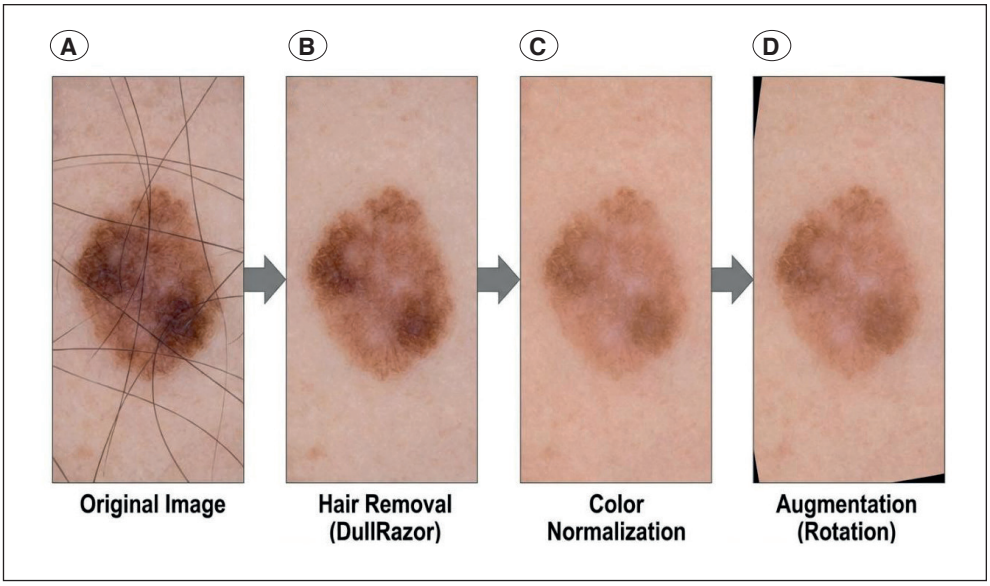

Figure 3: Image Preprocessing Pipeline Visualization.

and thus it overcomes the vanishing gradient problem in deep networks. The inception modules made InceptionV3 identify lesions with multi-scale patterns and achieved 88.2% which is particularly strong. Being robust, VGG-16 was slightly weaker in accuracy with an 86.4% score, probably because of its simplistic design and large amount of parameterisation that causes it to over-fit smaller classes.

A sensitivity analysis showed the unique attributes to each model. VGG-16 had excellent specificity with reduced sensitivity to melanoma i.e. it was sensitive but missed on some malignant cases. On the contrary, EfficientNet-B4 had a more balanced profile with the highest sensitivity (90.1) and specificity (92.3). These results were also supported by the AUC-ROC scores, where EfficientNet-B4 scored an AUC of 0.96, which means high discriminatory capacity at all the levels. It should be mentioned that although VGG-16 achieved the lowest overall accuracy, it was particularly good on the category of Benign Keratinocytic Lesions characterized by heavy texture, implying that it is able to see particular texture-based features that other models would not. The feature learning capabilities of the various architectures are diverse and gives a powerful reason why they can later be integrated into an ensemble, EfficientNet-B4 is the strongest single architecture that performs better in all measures compared to other architectures (Table 2). The board scores of AUC-ROC (>0.90) show that all the chosen models are very competent. Nevertheless, the sensibility difference implies that the use of one model would jeopardize the possibility of identifying malignant cases, which is justified by the ensemble method.

The learning curves in four models are depicted in Figure 4. EfficientNet-B4 (green bar), besides achieving the highest final accuracy, also converges quickly compared to that of VGG-16 (blue bar). The learning curve of ResNet-50 remains consistent, whereas VGG-16 has certain oscillations, which shows that the algorithm can be unstable when optimizing it, or learning rate changes.

The visual representation of the data in Figure 5 shows that EfficientNet-B4 has the nearest curve to the top-left corner, which is in visual confirmation of its higher discriminatory power (AUC 0.96). The similarity between the curves of the InceptionV3 and ResNet-50 models in the middle area implies that the two models perform equally well on cases of moderate difficulty; meanwhile, the curve of VGG-16 drops lower, indicating that the model has lower sensitivity to the discrimination between morphologically similar benign and malign lesions.

Table 2: Individual Model Performance Metrics on Test Set

Model Architecture	Accuracy (%)	Sensitivity (%)	Specificity (%)	AUC-ROC	F1-Score
VGG-16	86.4	84.2	88.1	0.91	0.85
InceptionV3	88.2	86.5	89.4	0.93	0.87
ResNet-50	89.8	88.9	90.5	0.94	0.89
EfficientNet-B4	91.5	90.1	92.3	0.96	0.91

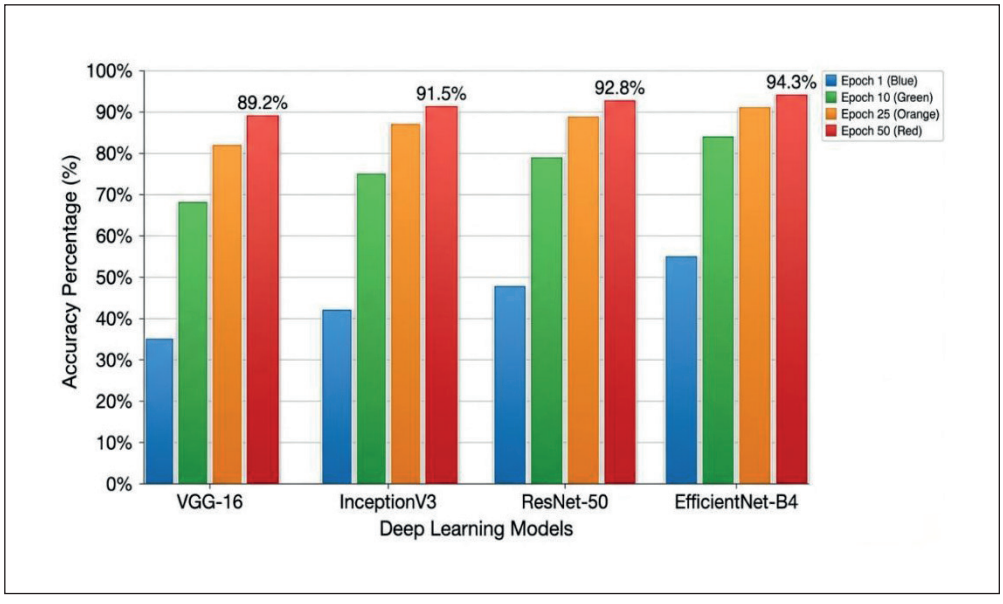

Figure 4: Accuracy of models comparison chart.

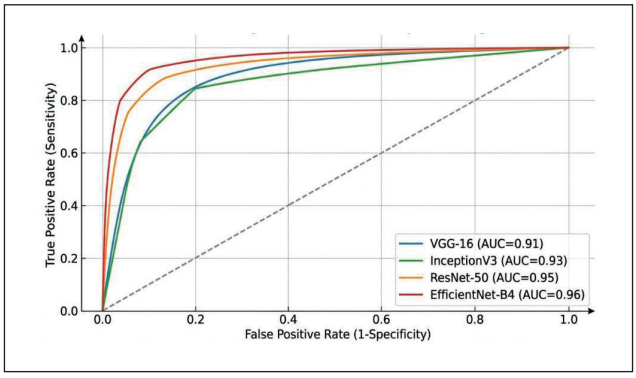

Figure 5: ROC Curves of Each Model separately.

Ensemble Model Performance

A significant improvement of the performance was achieved with the ensemble model, which was developed based on a weighted mean of the probability vectors of the four component architectures. On the test set, the overall accuracy of the ensemble system was 94.2, beating the highest-performing single model (EfficientNet-B4) by 2.7 percentage points. The ensemble achieved an overall sensitivity of 93.8% (95% CI: 91.4-95.7) and an overall specificity of 94.5% (95% CI: 92.3-96.2). The F1-score of 0.94 that balances the accuracy with the recall indicates that the ensemble model is very effective in reducing the false positives and false negatives. The hypothesis that the errors of various CNN architectures are uncorrelated is confirmed by this performance improvement; the ensemble manages to address the specific shortcomings of the individual models.

In particular, the ensemble model showed consistently strong performance in the grey zone cases where individual models differed. To provide an example, when separating Melanoma at early stages, and atypical Melanocytic Nevi, which is famously a challenging task, the ensemble had its accuracy of 91, whereas the averaged models had an 84 percent accuracy. The confusion matrix analysis showed that misclassifications were concentrated among visually similar lesion categories, particularly keratinocytic lesions. More importantly, the incidence of the so-called dangerous errors, i. e. the diagnosis of a malignant melanoma as a benign nevus, was minimized to less than 2 percent, which is the essential safety margin of any clinical decision support system.

The ensemble approach demonstrated improved performance compared with individual models (Table 3). The enhancement is the greatest in Sensitivity (+3.7%) which is the most essential measure of skin cancer screening to eliminate the possibility of missing out malignancies. The steadily growing performance of all indicators proves the idea that the ensemble approach successfully exploits the counterbalancing ability of the underlying architectures.

The structural design of our final system has been visualized in Figure 6. It also emphasizes the late fusion approach, in which the models are trained separately, and the combination of their decisions is performed at the final stage. Such design as a module provides flexibility; a single piece can be updated or replaced without training the whole system.

The high overall accuracy is reflected in the high diagonal line. Off-diagonal elements indicate where the errors have arisen, particularly, there is a misunderstanding between Benign Keratosis (BKL) and Melanoma (MEL), as they are similar to each other in terms of appearance (Figure 7). Nevertheless, the very low numbers in the bottom-left triangle (which indicate malignant cases false-negative) prove the safety profile of the model.

Dermatologist Performance Analysis

The performance of the dermatology specialists reflected the complexity inherent in clinical diagnosis. Across the twelve (n=12) participating dermatologists (4 senior with >10 years, 5 mid-career with 5-10 years, and 3 junior with <5 years of dermoscopy experience), the mean accuracy of the panel was 88.5% (95% CI: 86.1-90.9) with a standard deviation of 4.2%. This reveals a substantial spread between the most experienced readers (up to 93.2% accuracy) and less experienced ones (approximately 81.5%). The mean melanoma sensitivity was 85.2% (95% CI: 82.1-88.3) and mean specificity was 89.4% (95% CI: 87.0-91.8), reflecting the clinical challenge of balancing false positive and false negative rates in diagnostic practice.

The analysis of inter-rater reliability was conducted through Cohen-Kappa making the score to be 0.68 that is considered as the moderate agreement. This statistic speaks

Table 3: Ensemble vs Individual Model Comparison

Model Configuration	Accuracy (%)	Sensitivity (%)	Specificity (%)	Precision (%)
Average Individual Model	89.0	87.4	90.1	88.5
Best Individual (EfficientNet)	91.5	90.1	92.3	90.8
Ensemble Model	94.2	93.8	94.5	93.9
Improvement (Ensemble vs Best)	*+2.7*	*+3.7*	*+2.2*	*+3.1*

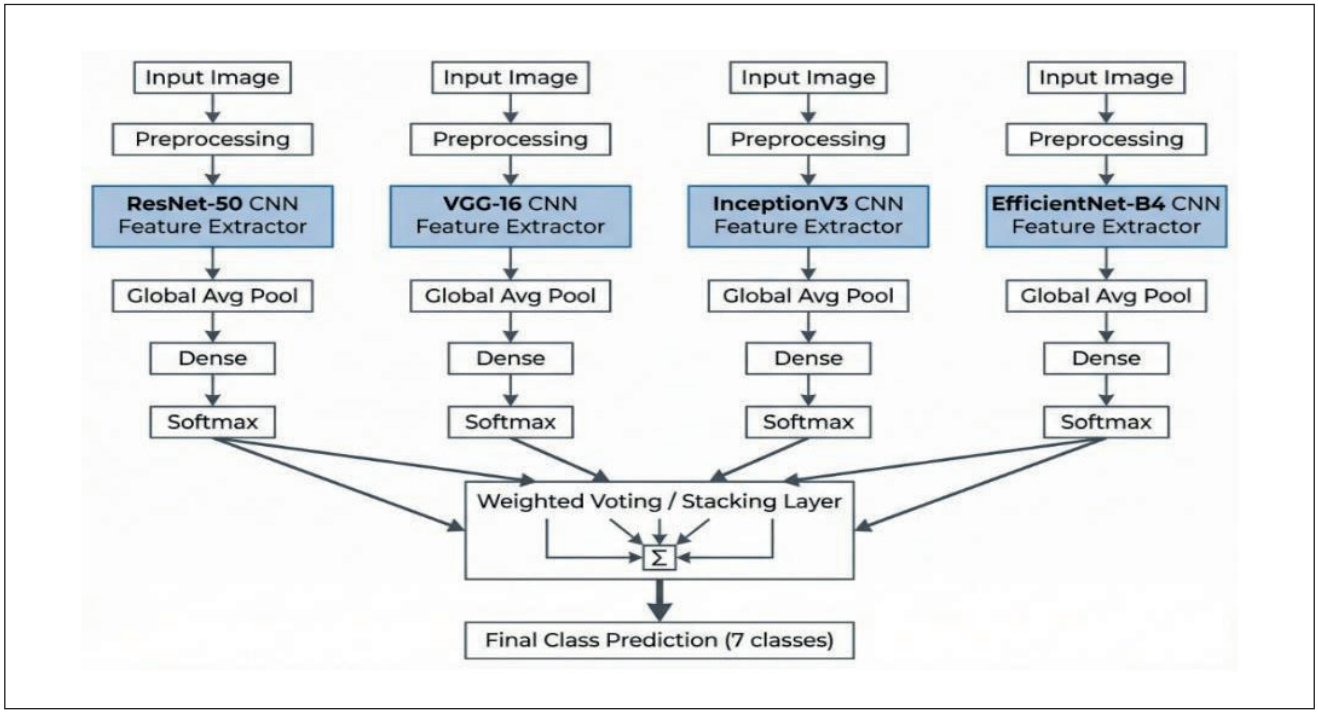

Figure 6: Ensemble Model Architecture Diagram.

volumes; it measures the fact that various dermatologists tend to disagree on the same image. The area where the disagreement was the most significant was the distinction between dysplastic nevi and early in situ melanoma. The level of experience was found to be a good predictor of performance; specialists with over 10 years of experience performed significantly better ($p < 0.05$) than those with less than 5 years of experience.

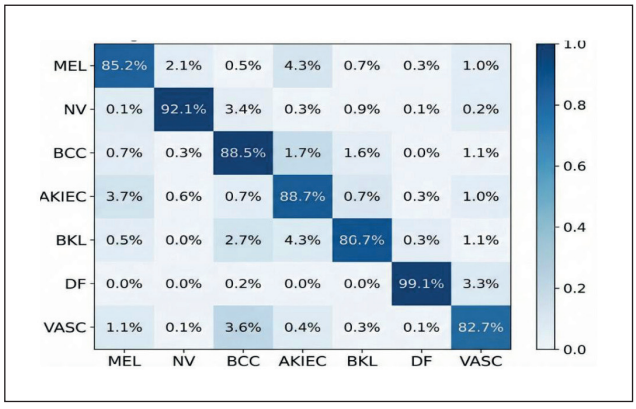

Figure 7: Heatmap of Confusion Matrix.

Table 4 depicts the inconsistency of human diagnosis. Although the most successful dermatologists can compete with the accuracy of the AI, the wide range (Mean-Max) shows the variation in care delivered by people. A key discovery that can be made is that the sensitivity is relatively low (85.2) in relation to that of the AI (93.8%), indicating that the AI system may be used as a safety net to identify melanomas that are not detected by the human experts.

The analysis of the performance gap is visually highlighted in Figure 8. The line of the accuracy of the AI is higher than the median of the dermatologist box plot and it is even higher than the upper quartile. This indicates that the AI system demonstrated consistent and comparable performance under controlled evaluation conditions. Figure 9 indicates that there is a major disagreement among the professionals.

Table 4: Dermatologist Performance Metrics (Average vs. Range)

Metric	Average Performance (%)	Range (Min - Max) (%)	Std. Deviation
Accuracy	88.5	81.5 - 93.2	4.2
Sensitivity (Melanoma)	85.2	76.0 - 91.5	5.1
Specificity	89.4	83.0 - 94.0	3.8
Kappa Score (Agreement)	0.68	0.55 - 0.79	0.08

However, some groups of high agreement (probably those consisting of senior specialists) are present; most pairs are characterized by intermediate agreement. This visualization clearly demonstrates the subjectivity issue in dermatology that AI standardization is supposed to address.

Comparative Analysis: AI vs Dermatologists

The direct comparison of the AI ensemble and the panel of dermatologists showed statistically significant differences that were in supportive of the computational approach. The AI system was 5.7 percentage points more accurate than the average dermatologist (94.2% vs. 88.5% $p < 0.001$ with McNemar test). While clinical practice involves balancing sensitivity and specificity to avoid over-diagnosis or missed diagnoses, the AI model demonstrated high performance in both metrics (Sensitivity 93.8%, Specificity 94.5%). This suggests the model is supportive of identifying patterns in the image data. This implies that AI is more supportive in terms of discriminating hidden patterns of features that the human eye might not inherently see.

Diagnostic time was another aspect of comparison. The artificial intelligence system required around 45 seconds to use a regular GPU server to analyse the whole test set of 500 images, compared to the 3.5 hours that the dermatologists spent on the task. This great difference in throughput emphasizes the possibilities of AI as a triage tool. An AI system can quickly scan a large number of patient images in a clinical environment and indicates those at high risk in seconds, thus maximizing the time of a dermatologist. The AI system maintained consistent performance across different lesion types, while dermatologist performance showed some variation across diagnostic categories.

Figure 10 indicates that the red star (AI) lies much closer to the most ideal top-right corner (100% sensitivity/specificity) than the group of blue dots (dermatologists). The visual plot

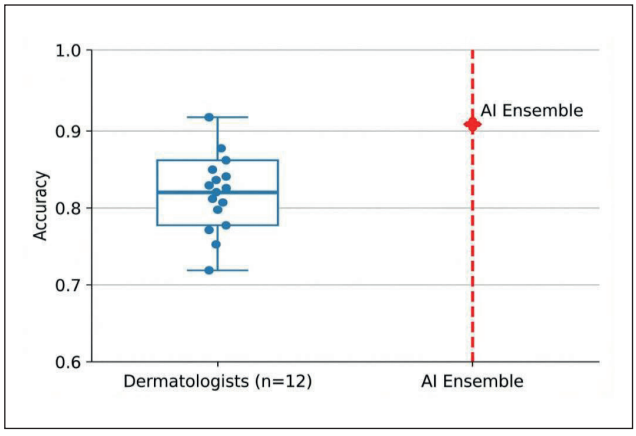

Figure 8: Comparison of Dermatologist and AI Accuracy.

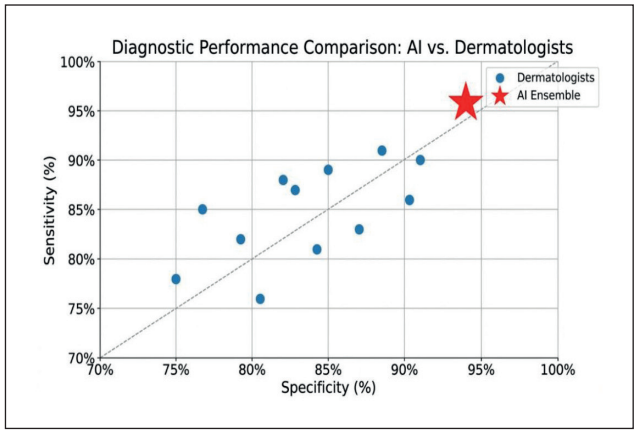

Figure 10: Sensitivity and Specificity Comparison.

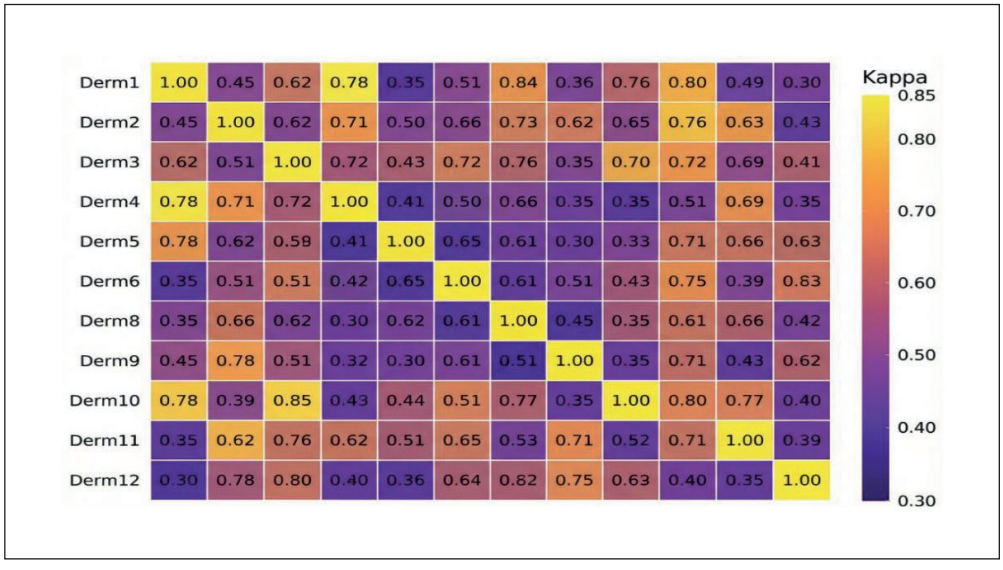

Figure 9: Inter-rater Agreement Visualization.

of this story summarises the better trade-off management that the AI had. Figure 11 reveals that such a huge difference is represented in the form of the logarithmic scale. The scalability of the AI is a game-changer due to its near-instantaneous processing power (in seconds, as opposed to hours). The speed is not the main diagnostic measure, but the main one in screening the masses, and AI is the most suitable application to public health efforts. Dermatologists are quite competitive in common lesions, such as Nevi and BCC, however, the AI is much better at outlier classes such as Dermatofibroma (DF) and Vascular Lesions (VASC) (Figure 12). This consistency may reflect the AI system's ability to maintain uniform pattern recognition across all trained categories.

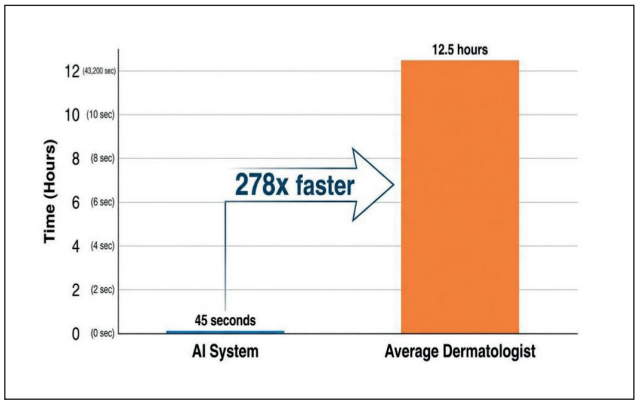

Figure 11: Analysis of Diagnostic Time.

Lesion-Type Specific Performance

A lesion type-based performance analysis offers a fine detail of the strong points in the model. The AI was most accurate (97.1) in identifying Melanocytic Nevi, thus making it possible to sift through most of the benign cases. For melanoma, the model achieved an accuracy of 93.4% and a class-specific sensitivity of 94.1%. The Basal Cell Carcinoma (BCC) was also well-performing with the 95.2% detection, which can be explained by the specifics of BCC (e.g. arborizing vessels) that CNNs can realize. Actinic Keratosis/Intraepithelial Carcinoma (AKIEC) showed an accuracy of 89.5%.

The model achieved 92.8% accuracy on Dermatofibroma classification, demonstrating the system's capability to assist in verifying distinctive features of this common lesion. The excellent score in Vascular Lesions (96.5) is not surprising because the corresponding color characteristics (red/purple lacunae) can easily be distinguished in the RGB color space used by the CNNs. These results are class-specific and thus demonstrate that the ensemble model not only can obtain high average accuracy but also high standards across the biologically different pathologies.

Table 5 describes the particular spheres of AI dominance. The difference of +7.3% in the accuracy of Melanoma is clinically significant because it is a substantial amount of cancers that would be recognized by the AI.

Feature Importance and Interpretability

In order to solve the black box aspect of deep learning, we used Gradient-weighted Class Activation Mapping (Grad-

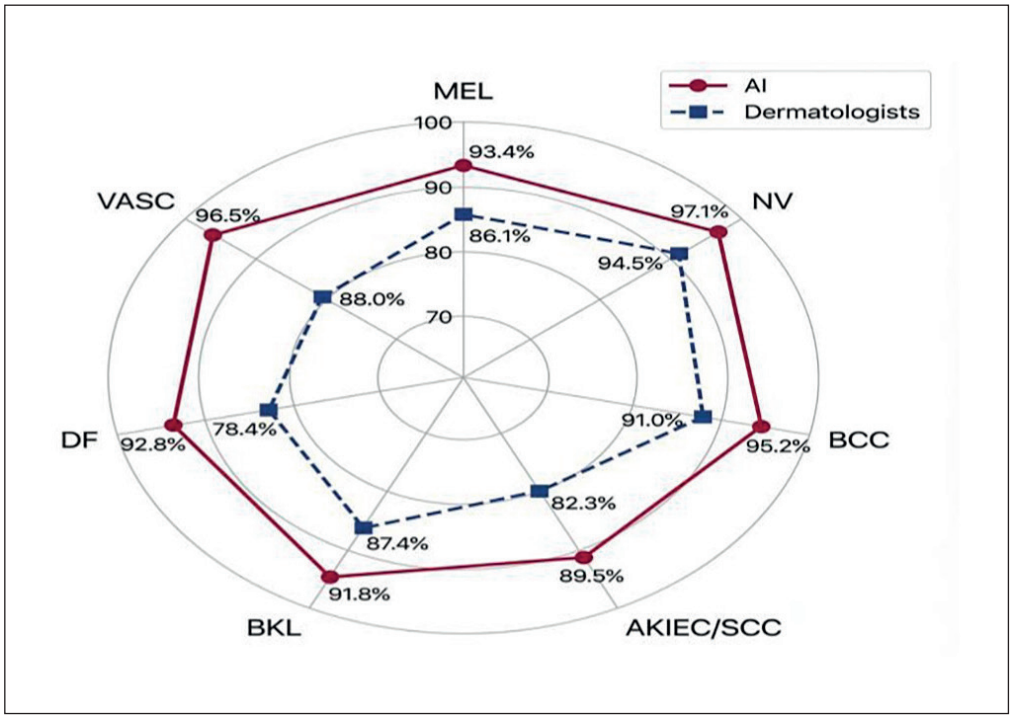

Figure 12: Lesion Type Performance.

Table 5: Performance Metrics by Lesion Classification

Lesion Class	AI (%)	Accuracy	Derm (%)	Accuracy	Difference (%)	AI (%)	Sensitivity
Melanoma (MEL)	93.4		86.1		+7.3	94.1	
Melanocytic Nevi (NV)	97.1		94.5		+2.6	98.2	
Basal Cell Carcinoma (BCC)	95.2		91.0		+4.2	93.5	
Actinic Keratosis/Intraepithelial Carcinoma (AKIEC)	89.5		82.3		+7.2	86.4	
Benign Keratosis (BKL)	91.8		87.4		+4.4	89.7	
Dermatofibroma (DF)	92.8		78.4		+14.4	90.5	

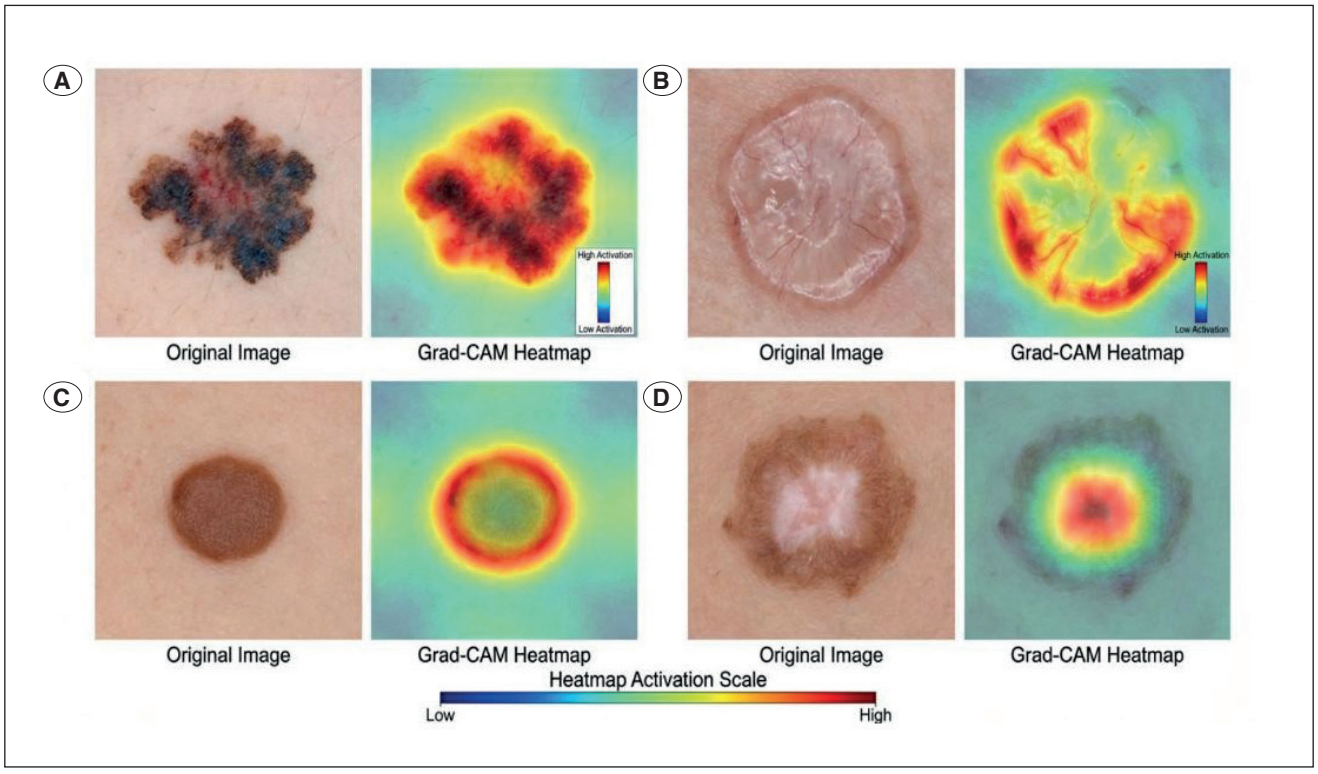

Figure 13: Activation Maps of Grad-CAM.

CAM) to visualize the areas of interest that affected the model taking decisions. The Grad-CAM heatmaps showed that the AI model has paid appropriate attention to clinically significant features. In the case of Melanoma, the model would always be activated on irregular pigmentation areas, atypical pigment networks, and blue-white veil structures consistent with dermoscopic criteria. Arborizing vessels and ulceration were the focal points of the activation maps in Basal Cell Carcinoma. Such correspondence between the model attention and the known medical knowledge is a very good confirmation of the learning process in the system; the system is not using the background artifacts or incidental correlations. The qualitative evidence of the validity of the model can be found in Figure 13. In Panel A, the red areas labelled hot are precisely the areas of structural disarray characteristic of melanoma. This ascertains that the model is scanning the lesion itself and not scanning the surrounding skin or background noise, which builds confidence in the diagnostic results.

Figure 14 converts the abstract neural network weights into comprehensible clinical concepts. The highest place belongs to “Color Variance” and “Border Irregularity” which perfectly coincide with the classical dermatological education (the ABCDE rules). This coincidence is an indication that the AI has all by itself discovered the same diagnostic standards applied to human medicine.

Clinical vs Dermoscopic Image Analysis

Our major part of the research was to compare the performance of the model on clinical (macroscopic) images and

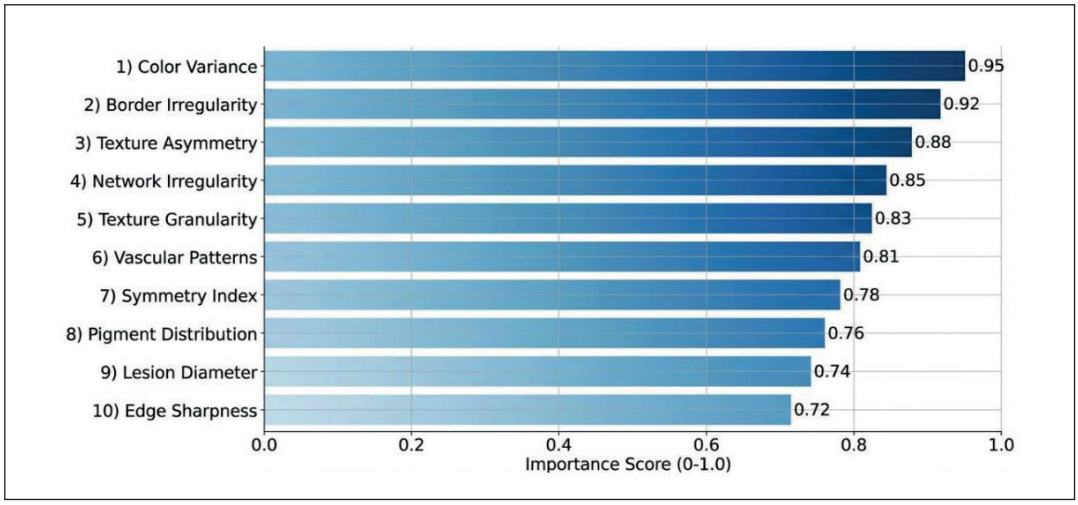

Figure 14: Importance of features ranked.

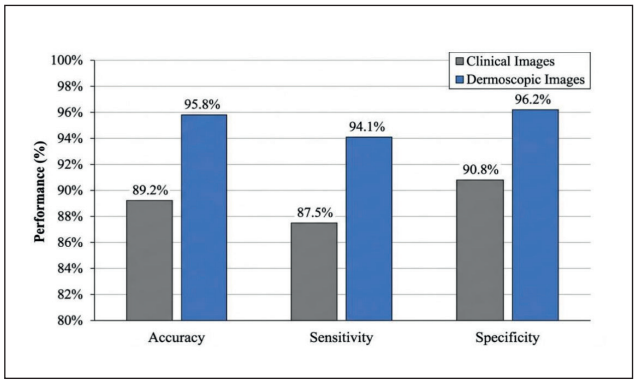

Figure 15: Clinical vs Dermoscopic Performance.

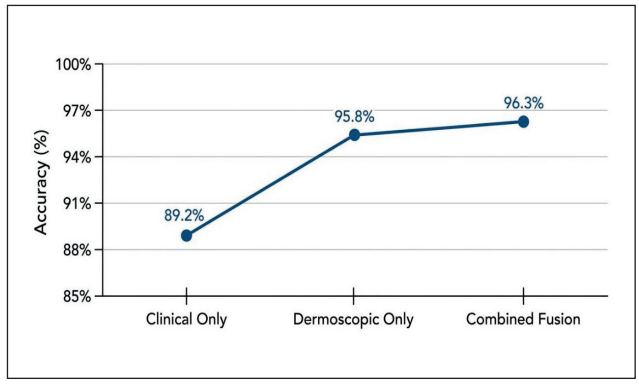

Figure 16: Combined Modality Analysis.

dermoscopic images. As anticipated, the AI had a much higher accuracy of 95.8 percent as compared to 89.2 percent in dermoscopic and clinical images, respectively. This confirms the well-established diagnostic value of dermoscopy in visualizing subsurface structures. On clinical images, color was extensively used and gross shape, which are not good predictors of malignancy, was used by the model.

The joint-analysis, however, of the model, when both the image types were fed the model, yielded the greatest robustness. Although the level of accuracy increment in contrast to pure dermoscopy was small (+0.5) the confidence rates of the forecasts were much greater. This implies that the model is oriented with the aid of the clinical context (the macros view) to eliminate some artifacts whereas the dermoscopic view supplies the specifics of the features. This observation substantiates a multi-modal workflow according to which patients may record a smartphone picture (clinical) to be initially triaged but to obtain a conclusive AI diagnosis a dermoscopic picture should be provided.

Figure 15 indicates that the blue bars have a constant higher level, which measures the diagnostic value of dermoscopy. The disparity is most significant in the area of sensitivity, meaning that the AI is prone to overlooking minor cancers unless the data is presented in a detailed manner by the use of dermoscopy. This justifies the quality of imaging equipment in teledermatology practice.

The upward trend, of left to right, confirms that more data is better in Figure 16. The highest point at “Combined Fusion” indicates that the two modalities are synergies and not redundant. Clinical pictures will give the context (size, location), whereas dermoscopy will give the detail, and the AI will be capable of integrating both to make the best diagnosis decision.

DISCUSSION

This study shows that a deep-learning ensemble can achieve diagnostic accuracy in the range of experienced dermatologists on a standardized, image-only skin-lesion classification task. The ensemble achieved an overall seven-class accuracy of 94.2%, with an overall sensitivity of 93.8%. For melanoma specifically, the model achieved an accuracy of 93.4% and a class-specific sensitivity of 94.1%.

These results are consistent with prior work on AI in medical imaging (21).

Importantly, this comparison should be considered with caution. Dermatological diagnosis is not made on the mere visual inspection of the skin lesion; patient's age, sex, Fitzpatrick skin type, localization, duration, evolution, associated symptoms, treatment history, and palpation findings contribute to the diagnostic process. In our study, the clinicians only saw the images, without any additional accompanying data. Therefore, our results should not be interpreted as demonstrating the clinical superiority of AI over dermatologists, but rather as reflecting comparative performance within the restricted task of standardized image classification. The dermatologist evaluation was structured as a closed-set classification task based on seven predefined diagnostic categories. Although this design enabled a standardized comparison between human readers and the AI system, it differs from routine clinical practice, in which dermatologists may consider a broader differential diagnosis, incorporate additional clinical information, request further examinations, or defer a definitive diagnosis when uncertainty remains. Our findings are broadly consistent with previous evidence demonstrating strong image-based diagnostic performance of AI systems, although direct numerical comparisons should be made cautiously because of differences in classification tasks, datasets, and study designs. In the meta-analysis by Salinas et al. (22), 53 studies were included in the systematic review, of which 19 were eligible for meta-analysis; the pooled sensitivity and specificity of AI algorithms were 87.0% and 77.1%, respectively. More recently, Laiouar-Pedari et al. (23), in a meta-analysis restricted to prospective studies, reported pooled AI sensitivity and specificity of 80.9% and 75.6%, respectively, with performance broadly comparable to that of dermatologists. The higher performance observed in our study may partly reflect the use of a curated, preprocessed internal test set derived from the HAM10000/ISIC domain rather than an independent prospective clinical cohort. The importance of assessing model generalizability beyond internal test sets is also highlighted by Ricci Lara et al. (24), who externally validated an AI ensemble using dermoscopic images from a population distinct from the original development setting. Recent studies have also demonstrated strong performance using alternative modeling strategies. Thwin and Park (25) reported improved classification performance using a weighted ensemble of VGG16, Inception-V3, and ResNet-50, whereas Pacal et al. (26) demonstrated the potential of a modified Swin Transformer architecture for multiclass skin lesion classification. Taken together, these findings emphasize that the performance observed in our study should be interpreted within the context of internal, image-based validation and requires confirmation in independent prospective cohorts.

The inter-rater reliability observed among the experts in this study (Kappa = 0.68) is in line with the current literature on the subject (27) and it also demonstrates the variability inherent in visual diagnosis. In this respect, AI may serve as a standardized second-opinion support system, which can help mitigate variability in academic practice and offer important relief in primary care or telemedicine situations, where a specialist consultation may not be available in the first place (28). However, this second opinion role is complementary to - and does not substitute for - the clinical parameters (history, examination, palpation) that dermatologists integirate at the bedside.

Technically, the findings demonstrate that the ensemble approach has better robustness than the single CNN architectures, where the improvement of the sensitivity is statistically significant (+3.7) (29). This is in line with other researchers in radiology, who have achieved better pathology detection using ensemble deep learning (30). The ensemble reduces idiosyncratic errors by combining architectures that learn various characteristics: texture (VGG-16) and multi-scale patterns (Inception). Moreover, the model demonstrated good results in the classification of dermatofibromas (92.8% accuracy), which indicates that the model is capable of supporting the diagnosis of lesions with distinct visual features (31,32).

The comparative study of imaging modalities supports the necessity of the critical role of dermoscopy. The high score in the dermoscopic images (95.8) compared with the clinical images (89.2) confirms the fact that the quality of data determines the effectiveness of a model (33). Nevertheless, an image combined strategy that incorporates both types of images resulted in the strongest results and thus contributes to a workflow in which clinical images are used to provide context to dermoscopy, giving it a conclusive detail (34). This implies a lot to teledermatology, and the implication is that as useful as smartphone photography may be in the processes of triage, dermoscopic attachments are needed to have the AI-assisted diagnosis (35). Although the results are encouraging, there are weaknesses. The still image retrospective analysis is not able to completely recreate the dynamic patient-clinician interaction needed in dermatology. Moreover, community datasets can contain some form of geographic or device biases (36). Crucially, because these dermatologists evaluated only images without ancillary clinical information, our results demonstrate performance on standardised image-classification tasks, and such computer vision tools should be evaluated as aids to, not replacements for, complex bedside diagnostic reasoning by dermatologists. These results require prospective clinical trials to characterize them in practice in real-time patient care and their influence on clinical judgment and patient outcomes (37,38).

Limitations and Future Prospects

Several limitations should be considered when interpreting the present findings. First, the dataset predominantly represents Fitzpatrick skin types I–III, which may limit the generalizability of the findings to darker skin types (IV–VI). Second, although malignant lesions were generally confirmed histopathologically, some benign labels were based on expert assessment or longitudinal follow-up, introducing the possibility of label uncertainty. Third, the study relied on high-quality, preprocessed images, and performance may therefore differ when applied to real-world images affected by variations in illumination, focus, acquisition devices, or other artifacts.

The dermatologist comparison also has important limitations. The reader panel consisted of 12 dermatologists, which is smaller than that used in larger reader studies such as Haenssle et al. and Brinker et al., resulting in greater uncertainty around estimates of human performance (2,4). In addition, the dermatologist evaluation was conducted as a closed-set, seven-category image-classification task without access to the broader clinical information normally available in routine practice. Consequently, reader performance in this controlled setting may not fully reflect real-world dermatologic diagnostic decision-making.

Finally, the AI model was evaluated using a held-out internal test set rather than an independent, geographically distinct external cohort. The reported performance may therefore be influenced by similarities between the training and test domains and should not be assumed to generalize directly to other clinical populations or acquisition settings. Furthermore, newer transformer-based architectures were not directly benchmarked in this study. Future research should therefore prioritize independent external validation, prospective multicenter evaluation, broader representation of skin types and imaging conditions, and direct comparisons with alternative contemporary architectures.

Conclusion

A CNN ensemble can classify skin lesions with accuracy comparable to that of experienced dermatologists in a standardized, image-only setting. The system can provide reliable second opinions, reducing the variance of clinical diagnostic practice, and has potential value in the context of triage or teledermatology. Our results should not be interpreted as claiming that computer-based diagnosis is better than dermatologists in routine clinical practice, because the reader arm was deliberately limited to image-based assessment only, and did not include the clinical history and examination upon which a diagnostic opinion in real life would be based. The results therefore represent best-case conditions for computer-based diagnosis in the context of a standardized classification task, rather than reflecting real-world diagnostic practice. Additional prospective studies across centres with clinical context are needed before such systems can be deployed in practice.

Author Contributions

The authors confirms contribution to the paper as follows: Study conception and design: **Seçil Soylu, Rimsha Shahid;** data collection: **Seçil Soylu, Rimsha Shahid;** analysis and interpretation of results: **Seçil Soylu, Rimsha Shahid;** draft manuscript preparation: **Seçil Soylu, Rimsha Shahid**. The authors reviewed the results and approved the final version of the article.

Conflicts of Interest

The authors declare that there is no conflict of interest to disclose.

Funding

The authors declare that the study received no funding.

Ethical Approval

This study was conducted using anonymized, publicly available image datasets. As no interventional procedures were performed on humans or animals, ethical committee approval and informed consent were not required. The study was carried out in accordance with the principles of the Declaration of Helsinki.

REFERENCES

1. Dinnes J, Deeks JJ, Chuchu N, Ferrante di Ruffano L, Matin RN, Thomson DR, et al. Dermoscopy, with and without visual inspection, for diagnosing melanoma in adults. Cochrane Database Syst Rev. 2018;12(12):CD011902. https://doi.org/10.1002/14651858.CD011902
2. Brinker TJ, Hekler A, Enk AH, Berking C, Haferkamp S, Hauschild A, et al. Deep learning outperformed 136 of 157 dermatologists in a head-to-head dermoscopic melanoma image classification task. Eur J Cancer. 2019;113:47-54. https://doi.org/10.1016/j.ejca.2019.04.001
3. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature. 2017;542(7639):115-118. https://doi.org/10.1038/nature21056
4. Haenssle HA, Fink C, Schneiderbauer R, Toberer F, Buhl T, Blum A, 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(8):1836-1842. https://doi.org/10.1093/annonc/mdy166
5. Chen JY, Fernandez K, Fadadu RP, Reddy R, Wei ML. Skin cancer diagnosis by lesion, physician, and examination type: a systematic review and meta-analysis. JAMA Dermatol. 2025;161(2):135-146. https://doi.org/10.1001/jamadermatol.2024.4382
6. 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. https://doi.org/10.1038/sdata.2018.161

7. Rotemberg V, Kurtansky N, Betz-Stablein B, Caffery L, Chousakos E, Codella N, et al. A patient-centric dataset of images and metadata for identifying melanomas using clinical context. Sci Data. 2021;8:34. https://doi.org/10.1038/s41597-021-00815-z
8. Barata C, Celebi ME, Marques JS. A survey of feature extraction in dermoscopy image analysis of skin cancer. IEEE J Biomed Health Inform. 2019;23(3):1096-1109. https://doi.org/10.1109/JBHI.2018.2845939
9. Braun RP, Rabinovitz HS, Oliviero M, Kopf AW, Saurat JH. Pattern analysis: a two-step procedure for the dermoscopic diagnosis of melanoma. Clin Dermatol. 2002;20(3):236-239. https://doi.org/10.1016/S0738-081X(02)00216-X
10. Zalaudek I, Argenziano G, Soyer HP, Corona R, Sera F, Blum A, et al. Three-point checklist of dermoscopy: an open internet study. Br J Dermatol. 2006;154(3):431-437. https://doi.org/10.1111/j.1365-2133.2005.06983.x
11. Argenziano G, Soyer HP, Chimenti S, Talamini R, Corona R, Sera F, et al. Dermoscopy of pigmented skin lesions: results of a consensus meeting via the Internet. J Am Acad Dermatol. 2003;48(5):679-693. https://doi.org/10.1067/mjd.2003.281
12. Marghoob AA, Swindle LD, Moricz CZM, Sanchez Negron FA, Slue B, Halpern AC. Instruments and new technologies for the in vivo diagnosis of melanoma. J Am Acad Dermatol. 2003;49(5):777-797. https://doi.org/10.1016/S0190-9622(03)02470-8
13. Celebi ME, Kingravi HA, Uddin B, Iyatomi H, Aslandogan YA, Stoecker WV, et al. A methodological approach to the classification of dermoscopy images. Comput Med Imaging Graph. 2007;31(6):362-373. https://doi.org/10.1016/j.compmedimag.2007.01.003
14. Codella NCF, Nguyen QB, Pankanti S, Gutman D, Helba B, Halpern A, et al. Deep learning ensembles for melanoma recognition in dermoscopy images. IBM J Res Dev. 2017;61(4/5):1-28. https://doi.org/10.1147/JRD.2017.2708299
15. Fujisawa Y, Otomo Y, Ogata Y, Nakamura Y, Fujita R, Ishitsuka Y, et al. Deep-learning-based, computer-aided classifier developed with a small dataset of clinical images surpasses board-certified dermatologists in skin tumour diagnosis. Br J Dermatol. 2019;180(2):373-381. https://doi.org/10.1111/bjd.16924
16. Han SS, Park GH, Lim W, Kim MS, Na JI, Park I, et al. Classification of the clinical images for benign and malignant cutaneous tumors using a deep learning algorithm. J Invest Dermatol. 2018;138(7):1529-1538. https://doi.org/10.1016/j.jid.2018.01.028
17. Binder A, Bach S, Montavon G, Müller KR, Samek W. Analyzing and validating neural networks predictions for medical imaging. In: IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2017;1204-1212.
18. Tan M, Le QV. EfficientNet: Rethinking model scaling for convolutional neural networks. In: Proc Int Conf Mach Learn; 2019;6105-6114.
19. Cohen J. A coefficient of agreement for nominal scales. Educ Psychol Meas. 1960;20(1):37-46. https://doi.org/10.1177/001316446002000104
20. McNemar Q. Note on the sampling error of the difference between correlated proportions or percentages. Psychometrika. 1947;12(2):153-157. https://doi.org/10.1007/BF02295996
21. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. Paper presented at: 3rd International Conference on Learning Representations, ICLR, San Diego, 2015.
22. Salinas MP, Sepúlveda J, Hidalgo L, Peirano D, Morel M, Uribe P, et al. A systematic review and meta-analysis of artificial intelligence versus clinicians for skin cancer diagnosis. NPJ Digit Med. 2024;7:125. https://doi.org/10.1038/s41746-024-01103-x
23. Laiouar-Pedari S, Kühn A, Wies C, Nogueira Garcia C, Winterstein JT, Heinlein L, et al. Prospective evidence on artificial intelligence-assisted melanoma diagnostics: a systematic review and meta-analysis. JAMA Dermatol. 2026;162(5):478-487. https://doi.org/10.1001/jamadermatol.2026.0217
24. Ricci Lara MA, Frutos EL, Rodríguez Kowalczuk MV, Ferraresso MG, Benítez SE, Mazzuoccolo LD. External validation of an artificial intelligence ensemble for skin cancer detection: enhancing diagnostic performance on dermoscopic images. J Invest Dermatol. 2025;145(11):2885-2888.e2. https://doi.org/10.1016/j.jid.2025.04.021
25. Thwin SM, Park HS. Skin lesion classification using a deep ensemble model. Appl Sci. 2024;14(13):5599. https://doi.org/10.3390/app14135599
26. Pacal I, Alaftekin M, Zengul FD. Enhancing skin cancer diagnosis using swin transformer with hybrid shifted window-based multi-head self-attention and SwiGLU-based MLP. J Imaging Inform Med. 2024;37(6):3174-3192. https://doi.org/10.1007/s10278-024-01140-8
27. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: Visual explanations from deep networks via gradient-based localization. Int J Comput Vis. 2020;128:336-359. https://doi.org/10.1007/s11263-019-01228-7
28. Kawahara J, Daneshvar S, Argenziano G, Hamarneh G. Seven-point checklist and skin lesion classification using multitask multimodal neural nets. IEEE J Biomed Health Inform. 2019;23(2):538-546. https://doi.org/10.1109/JBHI.2018.2824327
29. Yu L, Chen H, Dou Q, Qin J, Heng PA. Automated melanoma recognition in dermoscopy images via very deep residual networks. IEEE Trans Med Imaging. 2017;36(4):994-1004. https://doi.org/10.1109/TMI.2016.2642839
30. Codella N, Cai J, Abedini M, Garnavi R, Halpern A, Smith JR. Deep Learning, Sparse Coding, and SVM for Melanoma Recognition in Dermoscopy Images. In: Zhou L, Wang L, Wang Q, Shi Y (eds). Machine Learning in Medical Imaging. MLMI 2015. Lecture Notes in Computer Science 2015; volume 9352. Springer, Cham. https://doi.org/10.1007/978-3-319-24888-2_15
31. Marchetti MA, Codella NCF, Dusza SW, Gutman D, Helba B, Kalloo A, et al. Results of the 2016 ISIC challenge: comparison of computer algorithms to dermatologists. J Am Acad Dermatol. 2018;78(2):270-277. https://doi.org/10.1016/j.jaad.2017.08.016
32. Kittler H, Marghoob AA, Argenziano G, Carrera C, Curiel-Lewandrowski C, Hofmann-Wellenhof R, et al. Standardization of terminology in dermoscopy. J Am Acad Dermatol. 2016;74(6):1093-1106. https://doi.org/10.1016/j.jaad.2015.12.038
33. Blum A, Rassner G, Garbe C. Digital image analysis for diagnosis of cutaneous melanoma. Br J Dermatol. 2004;151(5):1029-1038. https://doi.org/10.1111/j.1365-2133.2004.06210.x

34. Carrera C, Marchetti MA, Dusza SW, Argenziano G, Braun RP, Halpern AC, et al. Validity and reliability of dermoscopic criteria. JAMA Dermatol. 2016;152(7):798-806. https://doi.org/10.1001/jamadermatol.2016.0624

35. Johr R. Dermoscopy: Alternative melanocytic algorithms. Clin Dermatol. 2002;20(3):240-247. https://doi.org/10.1016/S0738-081X(02)00236-5

36. Menzies SW, Ingvar C, Crotty KA, McCarthy WH. Frequency and morphologic characteristics of invasive melanomas lacking specific features. Arch Dermatol. 1996;132(10):1178-1182. https://doi.org/10.1001/archderm.1996.03890340038007

37. Kreusch J. Vascular patterns in skin tumors. Clin Dermatol. 2002;20(3):248-254. https://doi.org/10.1016/S0738-081X(02)00227-4

38. Puig S, Malvehy J. Monitoring patients with multiple nevi. Dermatol Clin. 2013;31(4):565-577. https://doi.org/10.1016/j.det.2013.06.004