

# Deep Learning-Based Multiclass Framework for Real-Time Melasma Severity Classification: Clinical Image Analysis and Model Interpretability Evaluation

Jun Zhang, Qian Jiang, Qiang Chen, Bin Hu, Liuqing Chen

Department of Dermatology, Wuhan No. 1 hospital, Wuhan, Hubei, People's Republic of China

Correspondence: Bin Hu; Liuqing Chen, Email binhu88@126.com; chlq35@126.com

**Background:** Melasma is a prevalent pigmentary disorder characterized by treatment resistance and high recurrence. Existing assessment methods like the Melasma Area and Severity Index (MASI) are subjective and prone to inter-observer variability.

**Objective:** This study aimed to develop an AI-assisted, real-time melasma severity classification framework based on deep learning and clinical facial images.

**Methods:** A total of 1368 anonymized facial images were collected from clinically diagnosed melasma patients. After image preprocessing and MASI-based labeling, six CNN architectures were trained and evaluated using PyTorch. Model performance was assessed through accuracy, precision, recall, F1-score, AUC, and interpretability via Layer-wise Relevance Propagation (LRP).

**Results:** GoogLeNet achieved the best performance, with an accuracy of 0.755 and an F1-score of 0.756. AUC values across severity levels reached 0.93 (mild), 0.86 (moderate), and 0.94 (severe). LRP analysis confirmed GoogLeNet's superior feature attribution.

**Conclusion:** This study presents a robust, interpretable deep learning model for melasma severity classification, offering enhanced diagnostic consistency. Future work will integrate multimodal data for more comprehensive assessment.

**Keywords:** melasma, deep learning, convolutional neural networks, MASI, clinical decision support

## Introduction

Melasma, a common hyperpigmentation disorder, predominantly affects middle-aged women, particularly those of Asian descent, with prevalence rates ranging from 8.8% to 40%.<sup>1</sup> Its pathogenesis is multifactorial, with key risk factors including UV exposure, hormonal therapies, and pregnancy.<sup>2</sup> Despite a variety of treatment modalities, melasma management remains challenging due to prolonged treatment durations, variable outcomes, and high recurrence rates.<sup>3</sup> Moreover, adherence to treatment is often compromised due to persistent pigmentation and high recurrence rates, significantly affecting patients' quality of life. These limitations underscore the need for early diagnosis, standardized monitoring tools, and enhanced patient education to optimize therapeutic outcomes.<sup>4</sup>

Current melasma assessment heavily relies on subjective scoring systems, such as the Melasma Area and Severity Index (MASI), which are influenced by clinician variability and patient compliance.<sup>5</sup> Additionally, patient dropout and inconsistent follow-up data hinder comprehensive evaluations of treatment efficacy, making it challenging to develop standardized treatment strategies. The lack of objective and quantitative assessment tools limits the ability to manage patient expectations effectively and optimize therapeutic approaches.<sup>6</sup> These limitations underscore the need for improved, technology-driven evaluation methods to enhance diagnostic accuracy and treatment monitoring.

In response to these challenges, deep learning and advanced imaging technologies are being explored to revolutionize melasma assessment.<sup>7,8</sup> Machine learning-based approaches offer the potential to provide objective, reproducible measurements of pigmentation depth and distribution, reducing reliance on subjective clinician evaluations. Our study

aims to develop a deep learning-based melasma assessment model trained on facial images, which could significantly improve diagnostic precision and disease monitoring. Notably, although all patients in this study underwent reflectance confocal microscopy (RCM) imaging, its detailed application in melasma assessment is not the focus of this research. Instead, we plan to integrate RCM-based analysis in future studies to further refine objective assessment methods and establish a comprehensive evaluation system.

By integrating artificial intelligence (AI) with clinical dermatology, this study seeks to bridge the gap between subjective evaluation and precise, data-driven analysis. Standardizing melasma assessment through AI-driven methodologies could significantly enhance early diagnosis, personalized treatment plans, and long-term patient outcomes. Collaborative efforts between dermatologists, researchers, and technology developers are crucial to validating these emerging techniques, ultimately transforming melasma management and improving the quality of life for affected individuals.

## Materials and Methods

### Design

This study included 1368 patients clinically diagnosed with melasma at the dermatology outpatient clinic of Wuhan First Hospital between January 1 and June 30, 2024. All participants underwent facial reflectance confocal microscopy (RCM) for diagnostic confirmation. The study focused exclusively on evaluating the area and depth of melasma lesions on the face.

To ensure high-quality and standardized image acquisition, all clinical photographs were captured using a Canon EOS 90D DSLR camera equipped with a 100mm f/2.8 macro lens. Uniform illumination was achieved using consistent LED ring lighting to minimize shadows and enhance clarity. Each image was taken at a fixed distance of 50 cm, with the camera positioned perpendicular to the patient's face to ensure consistency in angle and framing. A solid black background was used for all images to reduce visual distractions and enhance lesion contrast. Furthermore, all imaging sessions were conducted by the same trained medical photographer to minimize inter-operator variability.

Facial images were assessed using the Melasma Area and Severity Index (MASI) by three board-certified dermatologists, each performing an independent evaluation. To mitigate subjectivity and improve consistency, the final MASI score for each patient was calculated as the average of the three individual assessments. Severity was categorized based on the averaged score: mild ( $<16$ ), moderate (16–32), and severe (32–48). The labeled images were then randomly divided into training, validation, and testing datasets in a 6:3:1 ratio for model development and evaluation. A summary of the study workflow is shown in [Figure 1](#).

This study was conducted in accordance with the Declaration of Helsinki and was approved by the Institutional Review Board of Wuhan First Hospital (IRB No. [2023] 53). Written informed consent was obtained from all participants. To protect participant confidentiality, all personal identifiers were removed, and facial images were anonymized prior to analysis. The study adhered to all applicable national and international guidelines for human subject research.

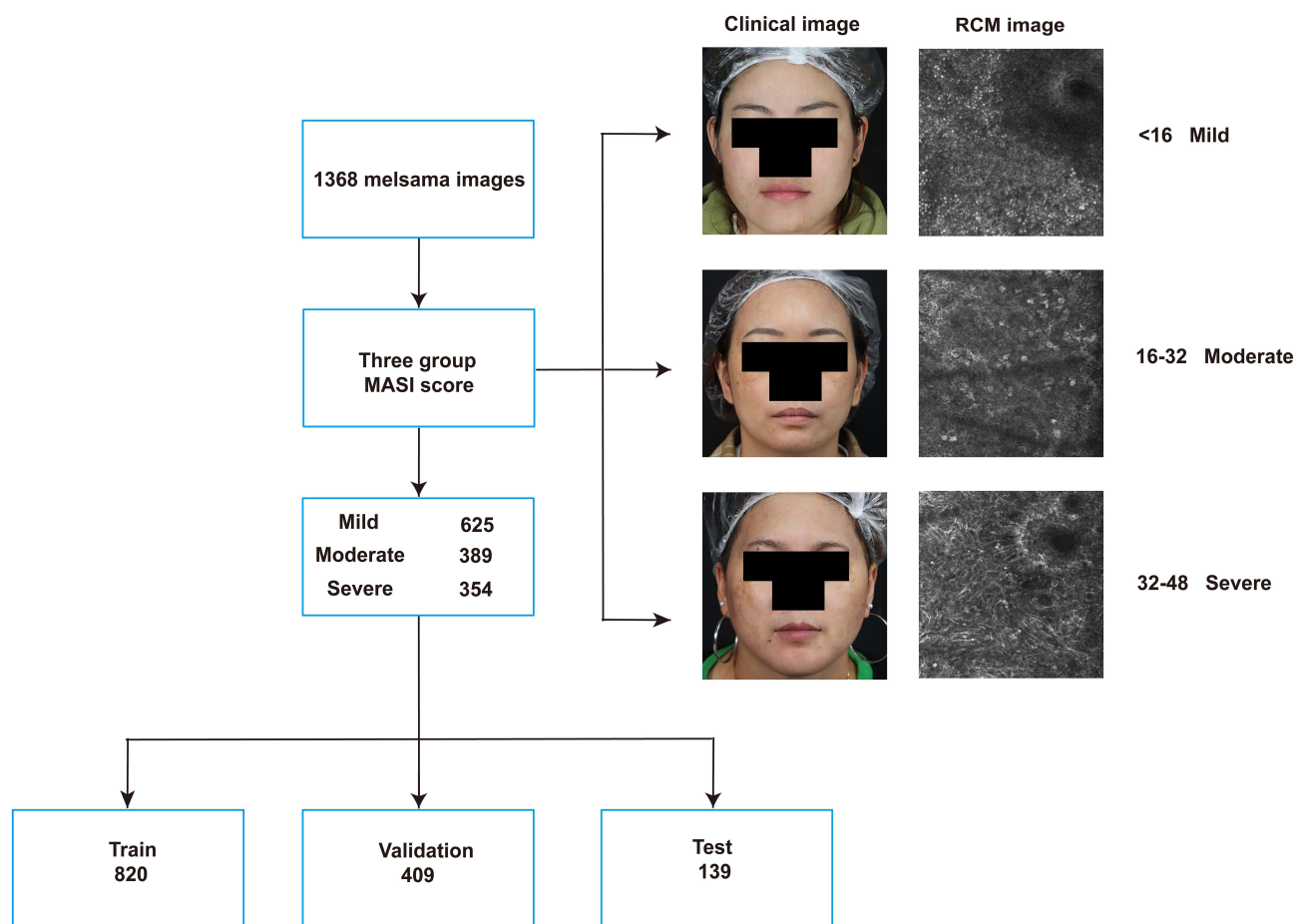
### Outcome Measures

PyTorch, as a leading deep learning framework, offers a variety of pre-trained deep learning models,<sup>1</sup> a widely adopted deep learning framework that provides access to a variety of pre-trained convolutional neural networks (CNNs). The pre-labeled facial images were used to train and evaluate six CNN architectures: AlexNet, ResNet-18, GoogLeNet, VGG16, DenseNet121, and MobileNetV2. All models were fine-tuned to classify melasma severity into three categories: mild, moderate, and severe.

Model performance was assessed using multiple standard classification metrics, including accuracy, precision, recall, F1-score, and area under the receiver operating characteristic curve (AUC). These metrics were computed using the following formulas:

#### True Positive (TP)

The model correctly predicts the positive class (both the prediction and actual values are positive).



**Figure 1** Overview of the study design and representative clinical and RCM images across melasma severity levels. The upper panel illustrates the study workflow, including patient recruitment, clinical image acquisition, preprocessing, MASI scoring, dataset partitioning, model training, and evaluation. The lower panel presents representative clinical photographs (left) and corresponding reflectance confocal microscopy (RCM) images (right) for mild, moderate, and severe melasma. Mild melasma is characterized by increased epidermal pigmentation without melanocyte hypertrophy. Moderate melasma shows enlarged and active melanocytes. Severe melasma exhibits highly reflective dendritic melanocytes, indicating intense pigment activity and distribution.

### True Negative (TN)

The model correctly predicts the negative class (both the prediction and actual values are negative).

### False Positive (FP)

The model incorrectly predicts the positive class when the actual class is negative (predicted-positive, actual-negative).

### False Negative (FN)

The model incorrectly predicts the negative class when the actual class is positive (predicted-negative, actual-positive).

The evaluation metrics are calculated as follows:

$$\text{Precision} = \frac{TP}{TP+FP} \quad \text{Recall} = \frac{TP}{TP+FN} \quad F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad \text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}$$

In this study, we used Layer-wise Relevance Propagation (LRP) to visualize and quantify model attention to melasma-related features. LRP is a widely used interpretability technique in medical AI and has been successfully applied in dermatology and histopathology to enhance model transparency and clinical trust.<sup>2-4</sup> Each model was initialized using PyTorch with pre-trained weights and adapted to a three-class output. Images were resized to 224×224 pixels and normalized before being passed through the network to obtain class predictions. LRP values were then computed by backpropagating the predicted class score and extracting relevance from the gradient-weighted input tensor. For each severity category, mean LRP scores were calculated across a validation set, and the results were statistically aggregated by model and severity level. These values were visualized using bar plots

to compare attribution patterns, and CSV files were generated for reproducibility. The LRP analysis provided insight into whether the models assigned higher relevance to image regions associated with increasing melasma severity, thereby supporting clinical interpretability and model transparency.

## Results

### Image Preprocessing and Annotation

A total of 1368 anonymized facial images were collected from patients clinically diagnosed with melasma at the Dermatology Department of Wuhan First Hospital. Prior to model training, all images underwent systematic preprocessing to enhance quality and ensure consistency across the dataset. Each image was resized to 224×224 pixels to match the input specifications of the convolutional neural networks (CNNs) used in this study. Normalization was applied to standardize pixel value distribution, improving training efficiency and model convergence.

To increase dataset diversity and prevent overfitting, data augmentation was performed using PyTorch's transformation pipeline. The augmentation techniques included random horizontal flipping, rotation ( $\pm 15^\circ$ ), and color jittering to simulate variations in lighting and skin tone. These transformations effectively enriched the training data and enhanced the generalization performance of the models.

For annotation, all images were independently evaluated by three board-certified dermatologists using the Melasma Area and Severity Index (MASI), a standardized scoring system for assessing melasma severity. Based on the averaged MASI scores, images were categorized into three severity levels: mild (MASI < 16), moderate (MASI 16–32), and severe (MASI 32–48).<sup>5</sup> A quality control protocol involving cross-verification among raters ensured annotation consistency.

In addition to MASI-based annotation, reflectance confocal microscopy (RCM) imaging was used to support severity classification and provide biological validation of melasma characteristics. RCM analysis revealed progressive changes in pigmentation and melanocyte morphology across severity levels. Mild melasma typically exhibited increased epidermal pigmentation without significant melanocyte alteration. In moderate cases, melanocytes appeared enlarged and more active, whereas severe melasma was characterized by highly active melanocytes with prominent, high-reflectivity dendritic structures. These observations were consistent with the MASI-based grading and further supported the reliability of the deep learning classification framework. Representative clinical photographs and corresponding RCM images for each severity level are shown in [Figure 1](#), illustrating the typical features observed in mild, moderate, and severe melasma.

### Model Development and Training

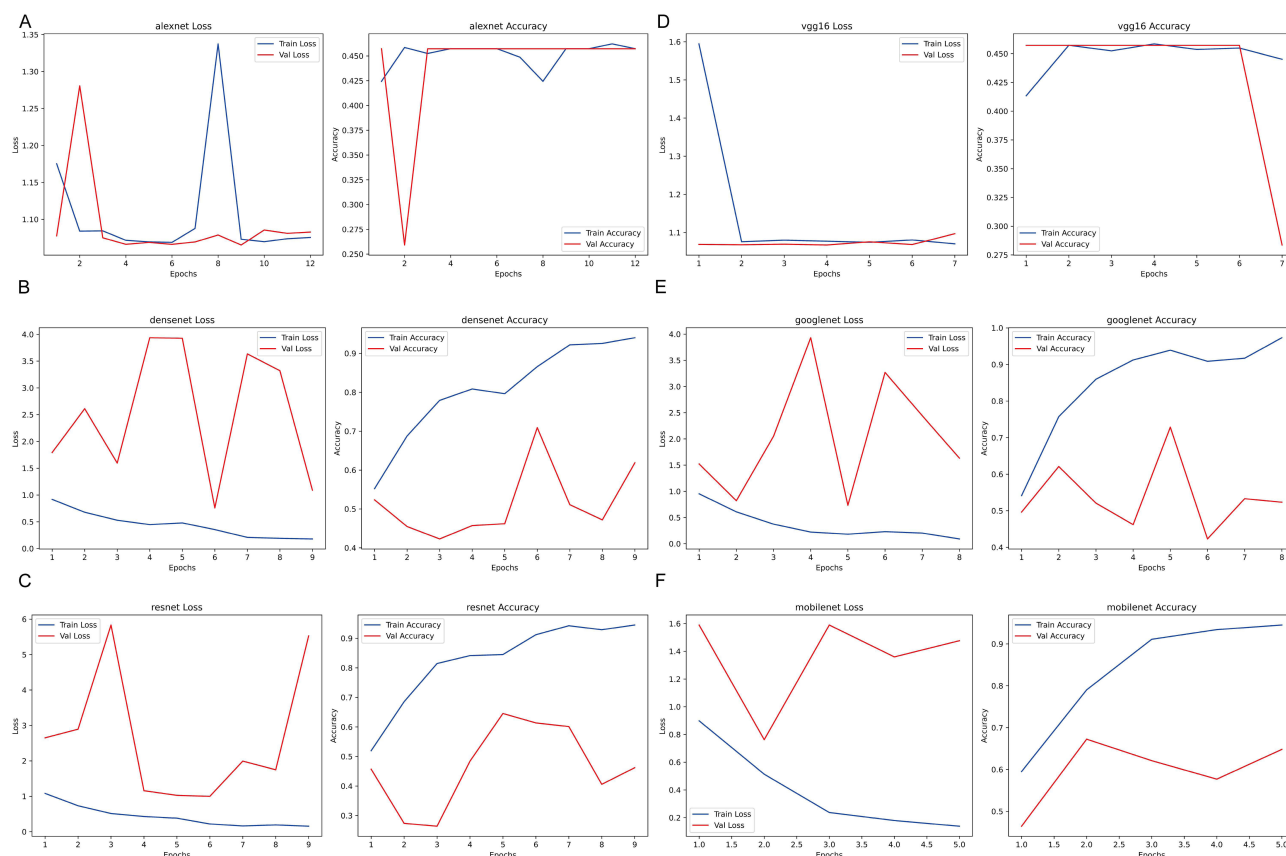
During the modeling phase, six widely used convolutional neural network (CNN) architectures—AlexNet, ResNet-18, GoogLeNet, VGG16, DenseNet121, and MobileNetV2—were selected from PyTorch's pre-trained model library and fine-tuned for classifying melasma severity.<sup>6–11</sup> Each model was adapted to accommodate three output classes (mild, moderate, severe), corresponding to MASI-graded and RCM-supported labels.

The training process utilized the Cross-Entropy Loss function to quantify the discrepancy between predicted and actual labels, while the Adam optimizer was employed for its efficiency in handling gradients, featuring a learning rate of 0.001 and a weight decay of 0.0001 to prevent overfitting.<sup>12,13</sup> Models were trained using a batch size of 32 over multiple epochs. Throughout the training cycle, training and validation losses and accuracies were continuously monitored and recorded.

To enhance generalization and reduce the risk of overfitting, an early stopping mechanism was implemented. Training was halted if validation loss failed to improve for three consecutive epochs. This approach ensured that model performance was optimized without excessive parameter updates. The training dynamics for each model, including learning curves and validation accuracy trends, are illustrated in [Figure 2](#), highlighting the convergence behavior and comparative learning stability across architectures.<sup>14</sup>

### Model Evaluation and Performance Comparison

Following model training, all six architectures were evaluated on a held-out test set using standard performance metrics, including accuracy, precision, recall, and F1-score.<sup>15</sup> Accuracy reflected the overall classification correctness, while



**Figure 2** Training and validation loss and accuracy curves for the six deep learning models. (A–F) present individual training process for each of the six convolutional neural networks (CNNs) evaluated: (A) AlexNet, (B) DenseNet, (C) ResNet, (D) VGG16, (E) GoogLeNet, and (F) MobileNet. GoogLeNet, DenseNet, and MobileNet demonstrated smoother convergence and lower validation loss across epochs, indicating improved generalization. In contrast, AlexNet and VGG16 exhibited unstable learning curves, consistent with their lower classification performance on the test set.

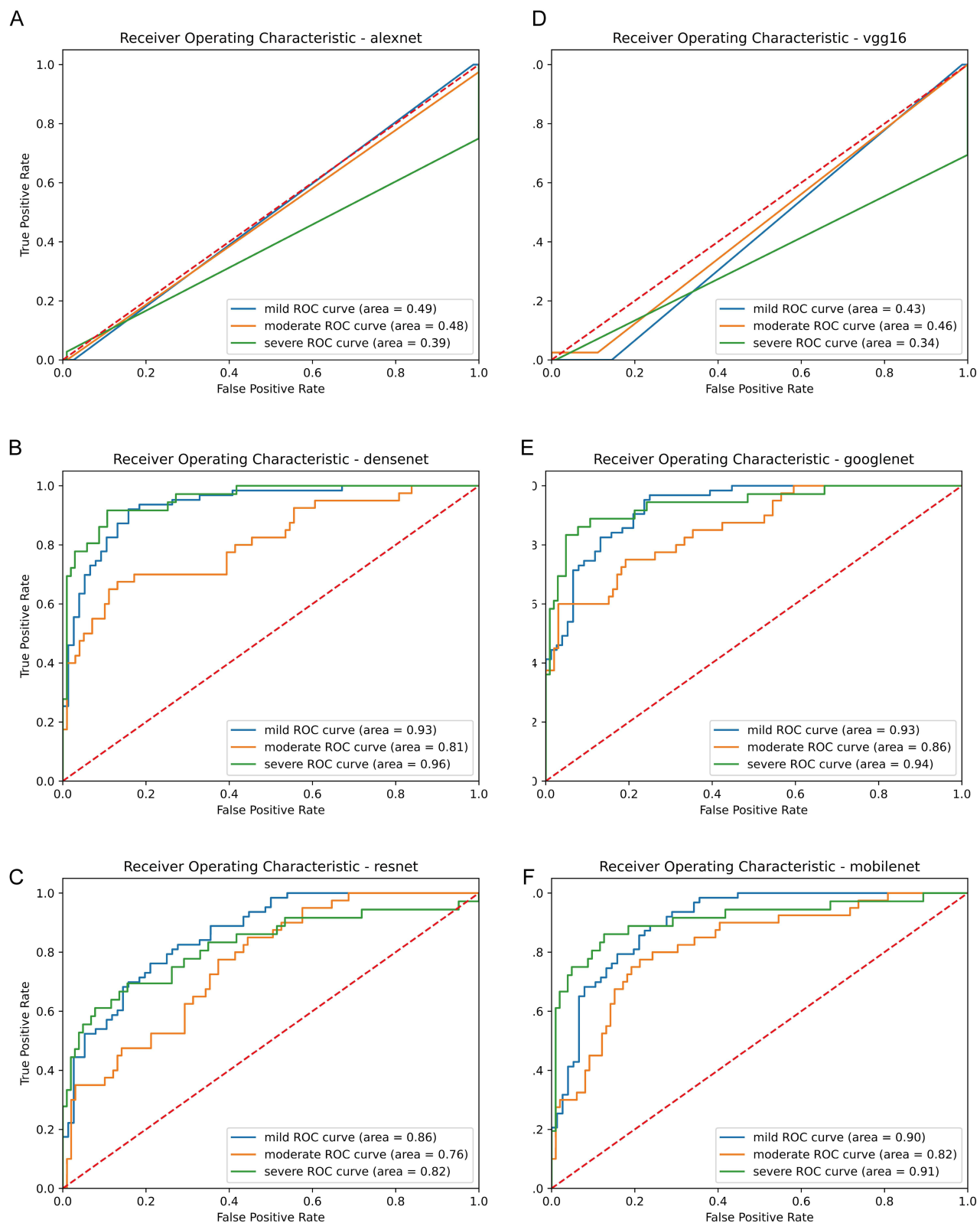
precision and recall assessed the model's ability to minimize false positives and capture true positives, respectively. The F1-score, as the harmonic mean of precision and recall, offered a balanced view of model performance. The quantitative results are summarized in Table 1.

Among the evaluated models, GoogLeNet outperformed all others, achieving the highest accuracy (0.755), precision (0.770), recall (0.755), and F1-score (0.756). DenseNet and MobileNet followed, with comparable accuracy scores of 0.727 and F1-scores of 0.712 and 0.694, respectively. In contrast, AlexNet and VGG16 performed poorly, with accuracy values of 0.453 and low F1-scores (0.283), indicating limited suitability for this classification task.

To further assess discriminative capacity, Receiver Operating Characteristic (ROC) curve analysis was performed for each model.<sup>16</sup> As shown in Figure 3, GoogLeNet achieved the highest AUC values across all three severity levels: 0.93 for mild, 0.86 for moderate, and 0.94 for severe melasma. These results indicate its superior ability to distinguish

**Table 1** Performance Metrics of Deep Learning Models for Melasma Severity Classification

Model	Accuracy	Precision	Recall	F1-Score
Alexnet	0.453237	0.205424	0.453237	0.282712
Resnet	0.625899	0.658339	0.453237	0.601982
Googlenet	0.755396	0.769569	0.755396	0.756077
Vgg16	0.453237	0.205424	0.453237	0.282712
Denseet	0.726619	0.783832	0.726619	0.711599
Mobilnet	0.726619	0.753674	0.726619	0.693812

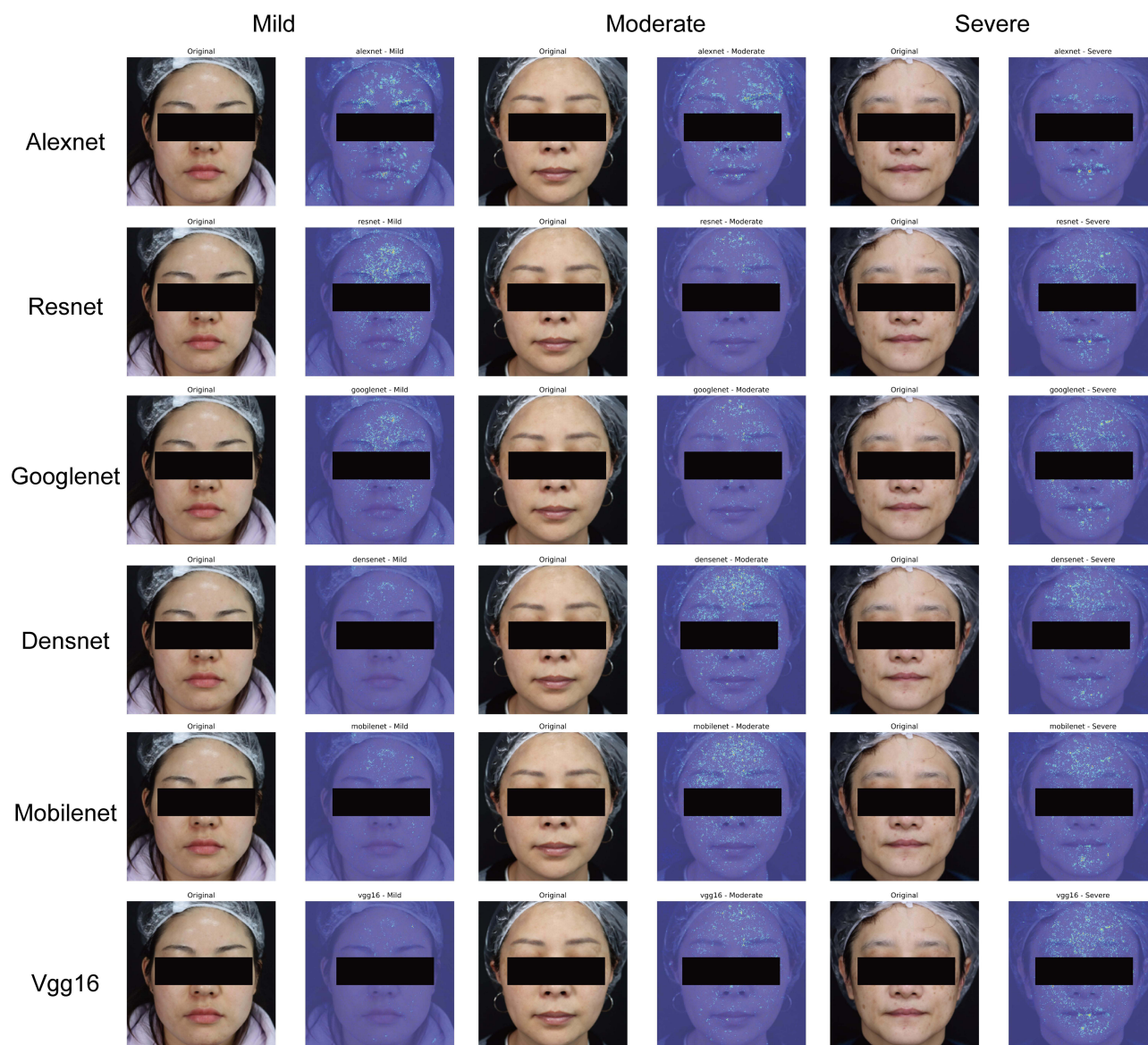


**Figure 3** Receiver Operating Characteristic (ROC) curves and AUC values for melasma severity classification. (A–F) present individual ROC curves for each of the six convolutional neural networks (CNNs) evaluated: (A) AlexNet, (B) ResNet, (C) GoogLeNet, (D) VGG16, (E) DenseNet, and (F) MobileNet. ROC curves were plotted for each model across mild, moderate, and severe classes. GoogLeNet achieved the highest AUC scores (0.93 for mild, 0.86 for moderate, and 0.94 for severe), indicating superior discriminative ability. ResNet showed moderate performance, while AlexNet and VGG16 had flatter curves, reflecting limited sensitivity and specificity.

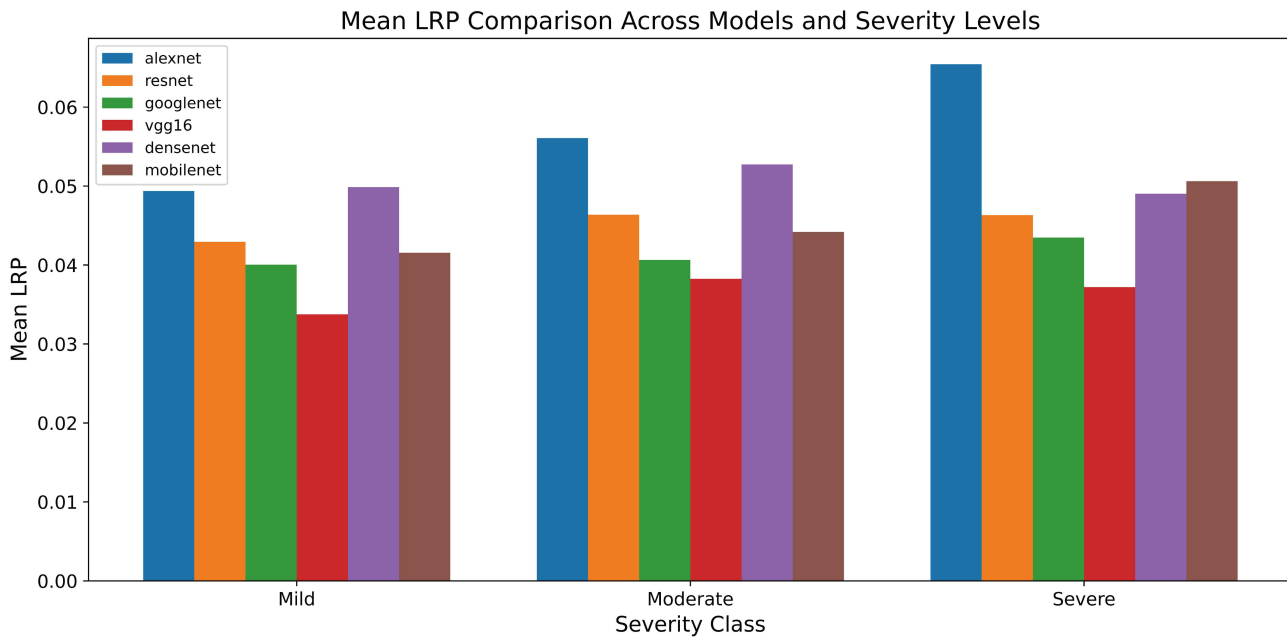


between clinically meaningful severity grades. ResNet followed with moderate AUC values, while AlexNet and VGG16 showed limited discriminative capability, consistent with their classification performance.

In addition to predictive metrics, Layer-wise Relevance Propagation (LRP) was employed to evaluate model interpretability. LRP quantified the average contribution of input image regions to the model's decision-making across severity categories. As illustrated in Figures 4 and 5 and Table 2, GoogLeNet demonstrated the lowest and most concentrated LRP values (Mild: 0.0351; Moderate: 0.0412; Severe: 0.0418), suggesting highly localized and efficient feature attribution. Conversely, AlexNet and VGG16 exhibited higher and more diffuse LRP values, indicating less precise decision focus. DenseNet and MobileNet offered moderate interpretability, balancing relevance intensity with spatial specificity. These patterns suggest that GoogLeNet not only achieved high classification accuracy but also aligned well with clinical patterns of pigmentation, enhancing trustworthiness in decision support.



**Figure 4** Layer-wise Relevance Propagation (LRP) heatmaps illustrating feature attribution for melasma severity classification. Representative heatmaps are shown for mild, moderate, and severe melasma cases across different models. The LRP visualizations highlight image regions contributing most strongly to the model's predictions. GoogLeNet displays more concentrated and interpretable activation patterns, whereas AlexNet and VGG16 exhibit more diffuse and less localized relevance signals, indicating reduced interpretability.



**Figure 5** Bar plots of mean LRP relevance values across models and melasma severity levels. This figure summarizes the average LRP scores for each CNN model when classifying mild, moderate, and severe melasma. Lower values indicate more focused and efficient feature attribution. GoogLeNet consistently demonstrates the lowest LRP scores across all severity levels, while AlexNet and VGG16 show higher, less specific attribution patterns.

Collectively, these findings validate GoogLeNet as the most accurate, robust, and interpretable model for melasma severity classification. Its superior performance across quantitative metrics, high AUC values, and focused LRP relevance patterns highlight its potential for integration into AI-assisted clinical decision-making. These results also underscore the importance of choosing appropriate architectures that balance performance and explainability when deploying deep learning tools in dermatological imaging tasks. While previous studies have focused primarily on binary classification for melasma detection or differentiating it from other hyperpigmented lesions, our study is among the first to implement a deep learning framework for severity stratification of melasma. By training on a MASI-labeled dataset and evaluating model interpretability via LRP, we provide a fine-grained, clinically relevant classification system that supports individualized management decisions.<sup>17</sup>

**Discussion**

**Deep Learning in the Clinical Assessment of Melasma: Opportunities and Challenges**

The utilization of deep learning models in the clinical assessment of melasma presents several transformative advantages, particularly in enhancing diagnostic accuracy and optimizing treatment protocols.<sup>18</sup> Melasma, characterized by its complex pathogenesis and variability in presentation, necessitates precise evaluation for effective management.<sup>19</sup> Traditional assessment methods, such as the Melasma Area and Severity Index (MASI), heavily depend on subjective evaluations from clinicians, which can lead to inconsistencies and biases in scoring.<sup>5</sup> In contrast, deep learning

**Table 2** Mean Layer-Wise Relevance Propagation (LRP) Values by Model and Severity Category

	AlexNet	ResNet	GoogLeNet	VGGI6	DenseNet	MobileNet
Mild	0.054575	0.046193	0.035092	0.034498	0.045951	0.042852
Moderate	0.060662	0.049674	0.041241	0.038392	0.054858	0.049603
Severe	0.060207	0.049125	0.041784	0.03666	0.055251	0.049445



methodologies can offer objective measurements, thereby reducing variability in assessments and enhancing the reliability of diagnoses.<sup>20</sup>

## Mechanistic Impact on Assessment and Management

### Objective Quantification of Pigmentation

These models can measure lesion area, intensity, and border sharpness quantitatively using image analysis, replacing subjective clinical judgment with pixel-based metrics.

### Region-Specific Analysis

Advanced models can localize pigmentation across facial regions (eg, cheeks, forehead, upper lip), offering more granular insights than traditional global severity scores. This helps in identifying lesion-prone areas for targeted interventions.

### Temporal Tracking

Longitudinal monitoring with deep learning enables precise tracking of changes in pigmentation over time. Clinicians can assess the efficacy of treatment regimens with consistent, automated scoring.

### Predictive Modeling

By integrating patient metadata (age, ethnicity, treatment history), deep learning models can potentially predict treatment response or relapse likelihood, supporting personalized treatment plans.

## Advantages of Pre-Classification Models

One of the primary benefits of deploying deep learning in melasma evaluation is the ability to analyze vast amounts of imaging data efficiently. Machine learning algorithms can identify patterns in pigmentation depth and distribution that may be imperceptible to the human eye.<sup>21</sup> This capability not only aids in the classification of melasma severity but also facilitates early diagnosis, which is crucial for initiating timely and appropriate treatment.<sup>22</sup> Early intervention is essential for preventing the worsening of pigmentation and improving overall patient outcomes.

Furthermore, pre-classification models streamline the diagnostic process by serving as an initial screening tool. They can categorize images into broad severity levels (mild, moderate, severe) quickly, allowing dermatologists to prioritize cases that may require immediate attention. This rapid triage capability enhances clinical workflows and improves patient management by enabling focused resource allocation.

## Limitations and Potential Biases

Despite these advantages, several challenges remain in the implementation of deep learning for melasma assessment.<sup>23</sup>

### Data Bias

Many deep learning models are trained on datasets that may not be representative of all skin tones, age groups, or ethnic backgrounds. This can lead to inaccurate assessments in underrepresented populations, perpetuating health disparities.

### Annotation Gaps

The absence of region-specific annotations limits the model's ability to understand pigment distribution nuances. Lesions on different facial zones may present distinct challenges requiring differentiated management strategies.

### Confounding with Other Conditions

Overlaps in visual features between melasma and other pigmented skin conditions (eg, post-inflammatory hyperpigmentation) can result in false positives or negatives, necessitating dermatologist oversight for verification.<sup>24</sup>

### Lack of Interpretability

Clinicians may be hesitant to adopt “black box” AI tools without understanding the rationale behind their predictions. Enhancing model explainability—through techniques like heatmaps or attention mechanisms—will be essential to gaining clinician trust.

## Performance and Applicability of Various Models

The deep learning models employed, while powerful, are not without their limitations. Each architecture has unique strengths and weaknesses that can impact performance in clinical settings:

### AlexNet

Simple and easy to implement, but prone to overfitting and limited in depth. Suitable for initial studies with limited data.<sup>6</sup>

### VGG16

High accuracy and effective for transfer learning, yet resource-intensive with a large model size. Works well in well-resourced environments.<sup>9</sup>

### GoogLeNet

Efficient with fewer parameters, allowing for multi-scale feature extraction, making it good for diverse manifestations of melasma. Its complexity may lead to longer training times but is beneficial for capturing various lesion characteristics.<sup>8</sup>

### ResNet

Extremely deep architecture that mitigates vanishing gradient issues, ideal for detailed classification with adequate resources. Its ability to learn complex features makes it particularly effective for accurate melasma assessments.<sup>7</sup>

### DenseNet

Efficient in parameter use and promotes feature reuse, making it strong for smaller datasets. However, it can be complex to train and has higher memory usage.<sup>10</sup>

### MobileNet

Lightweight and efficient, suitable for mobile applications, but may trade off some accuracy and robustness in complex datasets.<sup>11</sup>

The applicability of pre-classification models extends beyond initial assessments; they can be utilized in ongoing monitoring and follow-up care. By consistently evaluating the severity of melasma over time, these models can help track treatment efficacy and recurrence, facilitating timely adjustments to therapeutic strategies.

However, the challenges of model interpretability pose a significant barrier to clinical adoption. Clinicians may be hesitant to rely on “black box” models that offer little insight into the decision-making process behind predictions. To overcome this, ongoing research must focus on developing interpretability frameworks that elucidate how models arrive at specific classifications, fostering clinician trust in these technologies.

## Integrating Clinical Expertise

While the use of deep learning can enhance the accuracy of melasma assessments, it does not eliminate the need for clinical expertise. Dermatologists bring a wealth of knowledge regarding the multifactorial aspects of melasma, including hormonal influences, UV exposure, and individual patient history. Therefore, the integration of deep learning tools should complement, rather than replace, traditional clinical evaluations. A collaborative approach that combines the strengths of both technologies and human expertise will yield the best outcomes for patients.

## Comparison with Existing Literature

Recent advances in deep learning have demonstrated significant potential in skin lesion analysis and pigmented disorders, particularly in improving diagnostic accuracy and supporting clinical workflows. Tschandl et al showed that AI models could match or exceed dermatologist-level performance in skin cancer classification, highlighting the clinical applicability of CNNs in dermatological contexts.<sup>4</sup> While previous studies have focused primarily on binary classification for melasma detection or differentiating it from other hyperpigmented lesions, our study is among the first to implement a deep learning framework for severity stratification of melasma. By training on a MASI-labeled dataset and evaluating model interpretability via Layer-wise Relevance Propagation (LRP), we provide a fine-grained, clinically relevant classification system that supports individualized management decisions.<sup>17</sup>

## Future Directions

In terms of future directions, the evolution of deep learning methodologies holds great promise for advancing the management of melasma. Researchers should focus on enhancing dataset diversity to ensure that models generalize well across different demographics and skin types. This could involve expanding the cohort from which images are drawn, ensuring representation of various ethnic backgrounds and skin conditions. Additionally, exploring the integration of multimodal data—such as patient demographics, treatment history, and genetic factors—may yield deeper insights into the efficacy of treatments and help refine predictive models.

Continued collaboration among dermatologists, researchers, and technology developers is essential for addressing the current limitations and developing standardized protocols for melasma management. As these technologies advance, it will be crucial to ensure that they are implemented in ways that enhance clinical workflows, improve patient experiences, and ultimately lead to better health outcomes.

## Conclusion

This study demonstrates that deep learning provides a transformative solution for improving the accuracy, consistency, and clinical applicability of melasma severity assessment. By addressing the limitations of traditional methods such as subjective scoring and inter-observer variability, deep learning models offer objective, consistent, and data-driven tools that significantly enhance diagnostic accuracy and efficiency. Pre-classification models, in particular, demonstrate strong utility in streamlining clinical workflows and enabling early intervention through rapid image-based screening.

However, the successful implementation of these technologies depends on overcoming key challenges, including data bias, lack of region-specific annotations, and limited interpretability. Moreover, while AI systems can support clinical decision-making, they must not replace the nuanced insights of experienced dermatologists. A synergistic approach—combining technological innovation with clinical expertise—remains essential for optimal patient care.

Looking ahead, efforts should focus on diversifying training datasets, integrating multimodal data sources, and developing interpretable AI frameworks to foster trust and transparency. Through ongoing collaboration between clinicians, researchers, and engineers, deep learning technologies can be further refined and responsibly integrated into routine dermatological practice, ultimately improving patient outcomes in melasma care and beyond.

## Acknowledgments

This study was funded by the Innovation Group Project of the Hubei Provincial Department of Science and Technology (2022CFA037), “Precision Monitoring System for Biotargeted Therapy in Immune Skin Diseases”.

## Disclosure

The authors report no conflicts of interest in this work.

## References

1. Paszke A, Gross S, Massa F, et al. PyTorch: an imperative style, high-performance deep learning library. *arXiv preprint arXiv*. 2019
2. Bach S, Binder A, Montavon G, Klauschen F, Müller KR, Samek W. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLoS One*. 2015;10(7):e0130140. doi:10.1371/journal.pone.0130140
3. Binder A, Montavon G, Bach S, Müller KR, Samek W. Layer-wise relevance propagation for neural networks with local renormalization layers In: *Artificial Neural Networks and Machine Learning–ICANN 2016: 25th International Conference on Artificial Neural Networks, Barcelona, Spain, September 6–9, 2016, Proceedings, Part II* 25. Springer International Publishing; 2016. doi:10.48550/arXiv.1604.00825
4. Tschandl P, Rinner C, Apalla Z, et al. Human–computer collaboration for skin cancer recognition. *Nat Med*. 2020;26(8):1229–1234. doi:10.1038/s41591-020-0942-0
5. Rodrigues M, Ayala-Cortés AS, Rodríguez-Arámbula A, Hynan LS, Pandya AG. Interpretability of the modified melasma area and severity index (mMASI). *JAMA Dermatol*. 2016;152(9):1051–1052. doi:10.1001/jamadermatol.2016.1006
6. Xiao L, Yan Q, Deng S. Scene classification with improved AlexNet model. In: *2017 12th International Conference on Intelligent Systems and Knowledge Engineering (ISKE)*. 2017:1–6. doi:10.1109/ISKE.2017.8258820.
7. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016:770–778. doi:10.1109/CVPR.2016.90.
8. Szegedy C, Liu W, Jia Y, et al. Going deeper with convolutions. In: *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2015:1–9. doi:10.1109/CVPR.2015.7298594.

9. Mascarenhas S, Agarwal M. A comparison between VGG16, VGG19 and ResNet50 architecture frameworks for image classification. In: *2021 International Conference on Disruptive Technologies for Multi-Disciplinary Research and Applications (CENTCON). Vol. 1.* 2021:96–99. doi:10.1109/CENTCON52345.2021.9687944.
10. Huang G, Liu Z, Van Der Maaten L, Weinberger KQ. Densely connected convolutional networks. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR).* 2017:2261–2269. doi:10.1109/CVPR.2017.243.
11. Sinha D, El-Sharkawy M. Thin MobileNet: an enhanced MobileNet architecture. In: *2019 IEEE 10th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON).* 2019:0280–0285. doi:10.1109/UEMCON47517.2019.8993089.
12. Rezaei-Dastjerdehei MR, Mijani A, Fatemizadeh E. Addressing imbalance in multi-label classification using weighted cross entropy loss function. In: *2020 27th National and 5th International Iranian Conference on Biomedical Engineering (ICBME).* 2020:333–338. doi:10.1109/ICBME51989.2020.9319440.
13. Mehta S, Paunwala C, Vaidya B. CNN based traffic sign classification using Adam optimizer. In: *2019 International Conference on Intelligent Computing and Control Systems (ICCS).* 2019:1293–1298. doi:10.1109/ICCS45141.2019.9065537.
14. Shao Y, Taff GN, Walsh SJ. Comparison of early stopping criteria for neural-network-based subpixel classification. *IEEE Geosci Remote Sens Lett.* 2011;8(1):113–117. doi:10.1109/LGRS.2010.2052782
15. Dev S, Kumar B, Dobhal DC, Singh Negi H. Performance analysis and prediction of diabetes using various machine learning algorithms. In: *2022 4th International Conference on Advances in Computing, Communication Control and Networking (ICAC3N).* 2022:517–521. doi:10.1109/ICAC3N56670.2022.10074117.
16. Chang CI. An effective evaluation tool for hyperspectral target detection: 3D receiver operating characteristic curve analysis. *IEEE Trans Geosci Remote Sens.* 2021;59(6):5131–5153. doi:10.1109/TGRS.2020.3021671
17. Liu L, Liang C, Xue Y, et al. An intelligent diagnostic model for melasma based on deep learning and multimode image input. *Dermatol Ther.* 2023;13(2):569–579. doi:10.1007/s13555-022-00874-z
18. Chen X, Wang X, Zhang K, et al. Recent advances and clinical applications of deep learning in medical image analysis. *Med Image Anal.* 2022;79:102444. doi:10.1016/j.media.2022.102444
19. Passeron T, Picardo M. Melasma, a photoaging disorder. *Pigm Cell Melanoma Res.* 2018;31(4):461–465. doi:10.1111/pcmr.12684
20. van der Velden BHM, Kuijff HJ, Gilhuijs KGA, Viergever MA. Explainable artificial intelligence (XAI) in deep learning-based medical image analysis. *Med Image Anal.* 2022;79:102470. doi:10.1016/j.media.2022.102470
21. Deo RC. Machine learning in medicine. *Circulation.* 2015;132(20):1920–1930. doi:10.1161/CIRCULATIONAHA.115.001593
22. Jones OT, Matin RN, van der Schaar M, et al. Artificial intelligence and machine learning algorithms for early detection of skin cancer in community and primary care settings: a systematic review. *Lancet Digit Health.* 2022;4(6):e466–e476. doi:10.1016/S2589-7500(22)00023-1
23. Poldrack RA, Huckins G, Varoquaux G. Establishment of best practices for evidence for prediction: a review. *JAMA Psychiatry.* 2020;77(5):534–540. doi:10.1001/jamapsychiatry.2019.3671
24. Byun J, Kim D, Moon T. MAFA: managing False Negatives for Vision-Language Pre-training. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 2024:27314–27324. doi:10.48550/arXiv.2312.06112

## Clinical, Cosmetic and Investigational Dermatology

### Publish your work in this journal

Clinical, Cosmetic and Investigational Dermatology is an international, peer-reviewed, open access, online journal that focuses on the latest clinical and experimental research in all aspects of skin disease and cosmetic interventions. This journal is indexed on CAS. The manuscript management system is completely online and includes a very quick and fair peer-review system, which is all easy to use. Visit <http://www.dovepress.com/testimonials.php> to read real quotes from published authors.

Submit your manuscript here: <https://www.dovepress.com/clinical-cosmetic-and-investigational-dermatology-journal>

**Dovepress**  
Taylor & Francis Group