Instrumentation Mesure Métrologie
Vol. 25, No. 5, October, 2026, pp. 283-297
Journal homepage: http://iieta.org/journals/i2m

A Stacking Ensemble Framework for AI-Assisted Diagnostic Measurement of Melanoma from Dermoscopic Images

Brahim Akachat[1*], Nabil Messaoudi[1], Hamdane Mohamed Elkamel[2]

[1] ICOSI Laboratory, Abbes Laghrour University, Khenchela 40000, Algeria
[2] Teachers Training College of Constantine (ENSC) & LIRE Laboratory, University of Constantine 2, Constantine 25000, Algeria

Corresponding Author Email: akachat.brahim@univ-khenchela.dz

Copyright: ©2026 The authors. This article is published by IIETA and is licensed under the CC BY 4.0 license (http://creativecommons.org/licenses/by/4.0/).

https://doi.org/10.18280/i2m.250504

Received: 1 July 2026
Revised: 5 September 2026
Accepted: 14 September 2026
Available online: 31 October 2026

Keywords:
deep learning, transfer learning, dermoscopic imaging, stacking ensemble, diagnostic measurement, melanoma detection

ABSTRACT

Melanoma remains one of the most aggressive forms of skin cancer, and timely diagnosis is essential for improving survival outcomes. Recent advances in deep learning and transfer learning have substantially improved the automated classification of dermoscopic images, yet many single-model systems remain vulnerable to overfitting, instability, and limited generalization when trained on relatively small medical datasets. To address these challenges, this study proposes a robust stacking ensemble framework that combines DenseNet169, EfficientNetV2B0, and MobileNetV2 as base learners, with XGBoost serving as the meta-classifier. The architecture is designed to exploit the complementary strengths of lightweight convolutional neural networks (CNNs) while preserving computational efficiency and model reproducibility. The proposed workflow includes transfer learning, batch normalization, dropout regularization, standardized meta-feature construction, and ensemble learning. A review of the literature indicates that ensemble learning, threshold refinement, and transfer learning can each improve melanoma classification performance, but that these strategies are often implemented in isolation rather than in a unified clinical pipeline. Experimental results indicate that the proposed stacking strategy provides competitive classification performance and a favorable balance across key evaluation metrics. The selected G3 configuration, based on XGBoost with 40 training epochs, achieved 92.00% accuracy, 92.20% sensitivity, 91.80% specificity, 91.83% precision, 92.02% F1-score, and an area under the receiver operating characteristic curve (ROC-AUC) of 97.74% at the selected operating point. From an instrumentation and measurement perspective, the framework treats dermoscopic images as quantitative medical imaging data acquired through dermoscopic systems and provides an AI-assisted analysis layer for diagnostic measurement workflows. The framework is therefore positioned as a practical foundation for computer-aided melanoma detection, with potential value for clinical decision-support systems and future deployment in resource-conscious healthcare settings.

1. INTRODUCTION

Melanoma is one of the most aggressive forms of skin cancer because of its high metastatic potential and the strong dependence of patient survival on early diagnosis. Although significant advances have been achieved in surgical and immunological therapies, delayed detection remains a major cause of poor clinical outcomes. Consequently, early screening has become a central objective in dermatology, with dermoscopic imaging serving as a widely adopted non-invasive technique for examining suspicious skin lesions [1-3]. From a medical instrumentation and measurement perspective, dermoscopic images can be regarded as measurement data acquired by non-invasive dermoscopic imaging instruments. The proposed AI framework acts as an auxiliary analysis module that processes these measured image data and converts visual information into quantitative diagnostic indicators, such as sensitivity, specificity, and classification probabilities. In this sense, the system can be integrated into a computer-aided diagnostic measurement workflow to support clinicians in the assessment of suspicious skin lesions. Despite its clinical value, dermoscopic interpretation remains challenging because early-stage melanoma often shares visual characteristics with benign nevi, including similar color distributions, border irregularities, and structural patterns. Diagnostic performance therefore depends heavily on clinician expertise and may vary across observers, particularly in high-volume screening environments [4, 5].

Recent advances in artificial intelligence, particularly deep learning, have significantly improved automated medical image analysis. Convolutional neural networks (CNNs) automatically learn hierarchical image representations, eliminating the need for handcrafted feature extraction. Moreover, transfer learning enables knowledge acquired from

large-scale natural-image datasets to be effectively adapted to dermoscopic image analysis, making high-performance classification feasible even when medical datasets are relatively limited. Previous studies have also demonstrated that ensemble learning can enhance predictive robustness by combining complementary feature representations extracted by different CNN architectures [6-9].

Nevertheless, several practical challenges remain before these systems can be reliably adopted in clinical practice. Individual CNN models are often susceptible to overfitting, prediction instability, and sensitivity to variations in imaging conditions. Furthermore, many existing studies emphasize overall classification accuracy while paying limited attention to clinically important issues such as minimizing false-negative diagnoses, ensuring reproducibility, maintaining computational efficiency, and selecting decision thresholds that appropriately balance sensitivity and specificity. These limitations reduce the practical value of automated melanoma classification systems and hinder their adoption in real clinical environments [5, 10-12].

The growing demand for scalable diagnostic support further highlights the need for robust artificial intelligence solutions. Many healthcare systems face shortages of dermatology specialists, making timely assessment of suspicious skin lesions difficult. Consequently, artificial intelligence is increasingly viewed as a decision-support technology that complements rather than replaces clinicians by improving triage, supporting teledermatology, and facilitating large-scale screening programs [4, 11, 13].

Motivated by these clinical and technical challenges, this study addresses the need for a lightweight, robust, and clinically applicable melanoma classification framework. Unlike many existing approaches that rely on a single CNN architecture, fixed decision thresholds, or computationally demanding ensemble models, the proposed framework integrates transfer learning from three complementary pretrained CNNs—DenseNet169, EfficientNetV2B0, and MobileNetV2—through a stacking ensemble using an XGBoost meta-learner [14, 15]. In addition, clinically informed threshold optimization is employed to achieve a more appropriate balance between sensitivity and specificity, thereby reducing the risk of missed melanoma cases while preserving diagnostic reliability. By combining complementary feature representations with computational efficiency, reproducibility, and clinically relevant evaluation metrics, the proposed framework provides an effective decision-support solution for automated melanoma classification and demonstrates its potential for practical deployment in dermoscopic imaging workflows.

2. LITERATURE REVIEW

2.1 Deep learning for melanoma classification

Deep learning has substantially advanced automated melanoma and skin lesion classification, with CNNs demonstrating performance comparable to dermatologists in some image-based classification tasks [1]. Unlike traditional handcrafted approaches, CNNs can automatically learn discriminative representations from lesion morphology, color, and texture. Recent reviews, however, emphasize that reported performance depends not only on the architecture but also on dataset composition, preprocessing, class balance, and validation strategy [5, 12, 16]. Consequently, meaningful comparison between studies requires careful consideration of the complete experimental pipeline.

Recent research has also increasingly emphasized clinical relevance rather than benchmark performance alone. Prospective evaluations and systematic reviews highlight the gap between retrospective classification results and real-world clinical utility [4, 5]. For melanoma detection, this consideration is particularly important because false-negative predictions can delay treatment. Therefore, models should balance discriminative performance with sensitivity, computational efficiency, and clinically meaningful operating points.

2.2 Transfer learning and lightweight convolutional neural network architectures

Transfer learning is widely used in melanoma classification because pretrained networks can transfer representations learned from large datasets such as ImageNet to medical imaging tasks. This is particularly beneficial when annotated dermatological datasets are limited, reducing training requirements while providing effective initial feature representations. Transfer learning has been widely adopted for melanoma classification because pretrained CNNs can provide effective feature representations while reducing the training requirements associated with limited medical datasets [6, 8, 17].

Lightweight CNNs have also attracted attention because they provide a favorable balance between predictive performance and computational cost. MobileNetV2 is designed for computational efficiency, EfficientNetV2B0 provides efficient model scaling and training, and DenseNet169 promotes feature reuse through dense connectivity. Comparative studies have also demonstrated the effectiveness of transfer learning and lightweight deep-learning architectures for melanoma and skin-lesion classification [4, 8]. These findings motivate the use of diverse pretrained CNNs in the present study, where MobileNetV2, EfficientNetV2B0, and DenseNet169 are combined to exploit complementary representations while maintaining a manageable computational cost.

2.3 Ensemble learning and stacking strategies

Ensemble learning can improve medical image classification by combining complementary predictions from multiple models. Stacking extends conventional voting approaches by using a meta-learner to learn how the outputs of different base learners should be combined [14]. This is particularly relevant to melanoma classification because different CNN architectures may capture complementary characteristics of dermoscopic lesions.

Recent studies have reported strong performance from weighted and hybrid ensemble approaches on datasets such as International Skin Imaging Collaboration (ISIC) and HAM10000 [6, 7, 9, 18]. Other studies have demonstrated the potential of lightweight ensembles and combinations of deep learning with machine learning classifiers [9, 18]. Building on this literature, the present study employs DenseNet169, EfficientNetV2B0, and MobileNetV2 as complementary base learners and uses XGBoost as the meta-classifier. This design aims to exploit differences among CNN predictions while maintaining a computationally manageable stacking

architecture.

Previous studies have indicated that model diversity can improve predictive performance, although computational cost and interpretability remain important considerations [12, 18]. The proposed framework addresses these considerations by limiting the ensemble to three pretrained CNNs and using a lightweight meta-learning stage.

2.4 Threshold optimization and diagnostic balance

Accuracy alone is insufficient for evaluating medical diagnostic systems because false negatives and false positives have different clinical consequences. In melanoma detection, sensitivity is particularly important because missed malignant lesions may delay treatment, while specificity remains relevant for limiting unnecessary clinical reviews. Threshold optimization therefore provides an important mechanism for converting probabilistic model outputs into clinically meaningful decisions.

ROC-based approaches, including the Youden index, provide a principled method for selecting an operating point that balances sensitivity and specificity [19]. Recent medical AI studies emphasize that model selection should consider clinically relevant error profiles rather than relying exclusively on overall accuracy [10]. Accordingly, the present study incorporates sensitivity-oriented threshold optimization into the decision pipeline, allowing the operating point to reflect the priority of detecting malignant lesions.

2.5 Research gap and motivation

Despite substantial progress in transfer learning and ensemble-based melanoma classification, several limitations remain. Many existing approaches rely on individual pretrained architectures, while others employ ensemble strategies that may increase computational complexity [6, 7, 9, 12, 17, 18]. Moreover, systematic reviews continue to identify external validation, explainability, and clinical translation as important challenges for melanoma AI systems [5, 16].

These limitations motivate a framework that combines architectural diversity with computational efficiency and clinically oriented decision optimization. The present study therefore integrates three complementary pretrained CNNs—MobileNetV2, EfficientNetV2B0, and DenseNet169—with an XGBoost meta-learner and sensitivity-oriented threshold optimization. This combination is intended to exploit complementary image representations while maintaining a relatively compact and reproducible pipeline. The framework thus addresses the identified gap by moving beyond single-model classification toward a stacking-based approach that considers both predictive performance and the clinical importance of the selected operating point.

3. CONCEPTUAL FRAMEWORK

Figure 1 illustrates the framework. It positions transfer learning-based deep neural networks as the principal predictive engine for melanoma detection, with ensemble stacking serving as the mechanism through which the outputs of multiple models are integrated. The independent variables are the three pretrained CNN architectures—MobileNetV2, EfficientNetV2B0, and DenseNet169—together with the XGBoost meta-learner. The dependent variable is melanoma classification performance, reflected through accuracy, sensitivity, specificity, precision, F1-score, and area under the receiver operating characteristic curve (ROC-AUC). The framework assumes that model diversity improves robustness and that a meta-learner can exploit complementary feature representations more effectively than any single backbone alone.

The model is grounded in the assumption that melanoma detection benefits from both architectural diversity and clinically informed decision selection. MobileNetV2 contributes computational efficiency, EfficientNetV2B0 contributes balanced scaling, and DenseNet169 contributes dense feature reuse. When these outputs are combined through a meta-learner, with XGBoost selected as the final meta-classifier, the resulting system is expected to exploit complementary information from the base learners. Such a system should reduce variance and improve robustness. This conceptual structure is consistent with established principles of ensemble learning [14] and clinical decision support [4, 11].

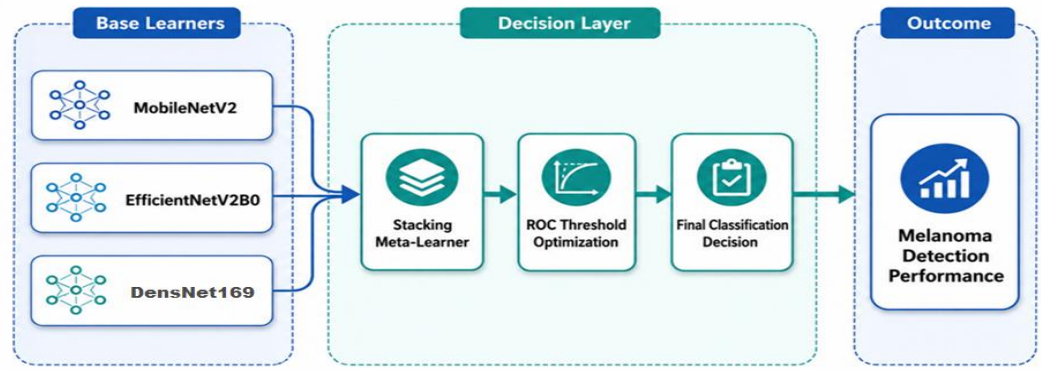


Figure 1. Conceptual framework diagram

4. METHODOLOGY

4.1 Study design

The study adopts a comparative experimental design to evaluate a stacking ensemble for binary melanoma classification. The proposed framework employs three pretrained CNNs, namely MobileNetV2, EfficientNetV2B0, and DenseNet169, as base learners. Their predictions are subsequently combined using stacked generalization [14],

with XGBoost serving as the meta-classifier to model nonlinear relationships among the base-learner outputs [15]. The overall architecture of the proposed stacking framework for melanoma detection is illustrated in Figure 2.

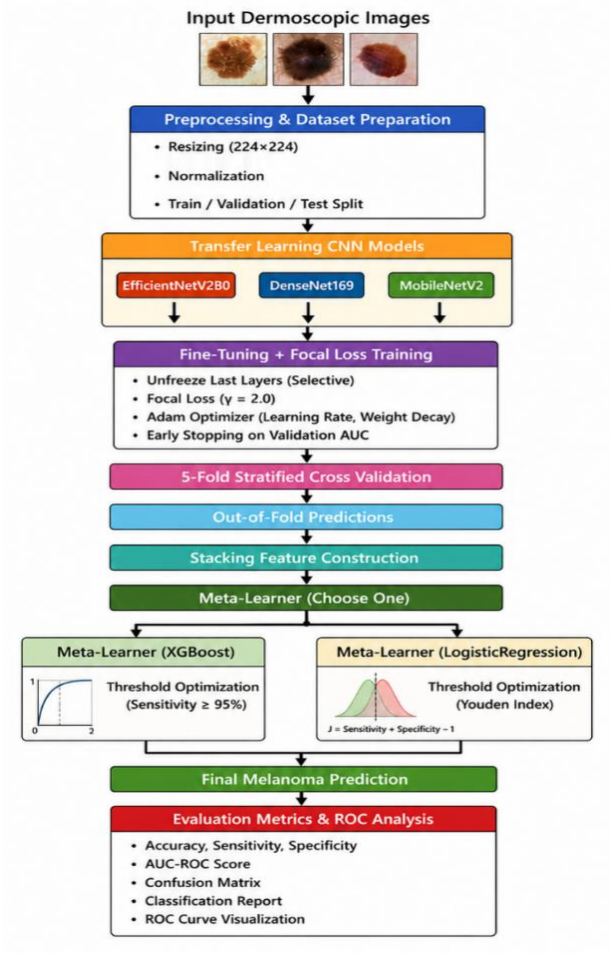


Figure 2. Conceptual framework of the proposed stacking ensemble for clinical melanoma detection

This design is appropriate because the research objective is not to develop a novel segmentation procedure, but to test whether ensemble fusion of complementary feature extractors improves classification robustness and generalization. Transfer learning is used to reduce training burden and adapt the pretrained feature spaces to dermoscopic imagery, which is consistent with prior work in medical image analysis [6, 20].

The overall logic of the design is grounded in ensemble theory and clinical decision support. Rather than relying on one architecture to learn every diagnostic cue, the framework allows several networks to specialize differently and then combines their outputs through a second-stage learner. This is advantageous in melanoma detection, where lesion appearance can vary substantially across cases and where no single architecture is guaranteed to dominate across all subsets of images [7, 18]. The method therefore seeks both statistical diversity and operational stability.

4.2 Dataset and preprocessing

The experiments were conducted using the publicly available Kaggle Melanoma Skin Cancer Dataset [21], which was compiled from publicly available skin-lesion repositories, primarily the ISIC Archive and the HAM10000 dataset [22]. It comprises 9,605 images for model development and 1,000 test images for final performance evaluation. The 9,605 development images were divided into training (80%) and validation (20%) subsets, while the test set was kept separate and reserved exclusively for final evaluation. The training and validation subsets showed slight class imbalance, whereas the test set was balanced between benign and malignant lesions. The corresponding data distribution is summarized in Table 1, and representative benign and malignant samples from the test set are shown in Figure 3. During preprocessing, all images were resized to 224 × 224 pixels, represented in the RGB color space to ensure compatibility with the pretrained CNN architectures, and normalized prior to training. In addition, class weights were applied to mitigate residual class imbalance during fine-tuning.

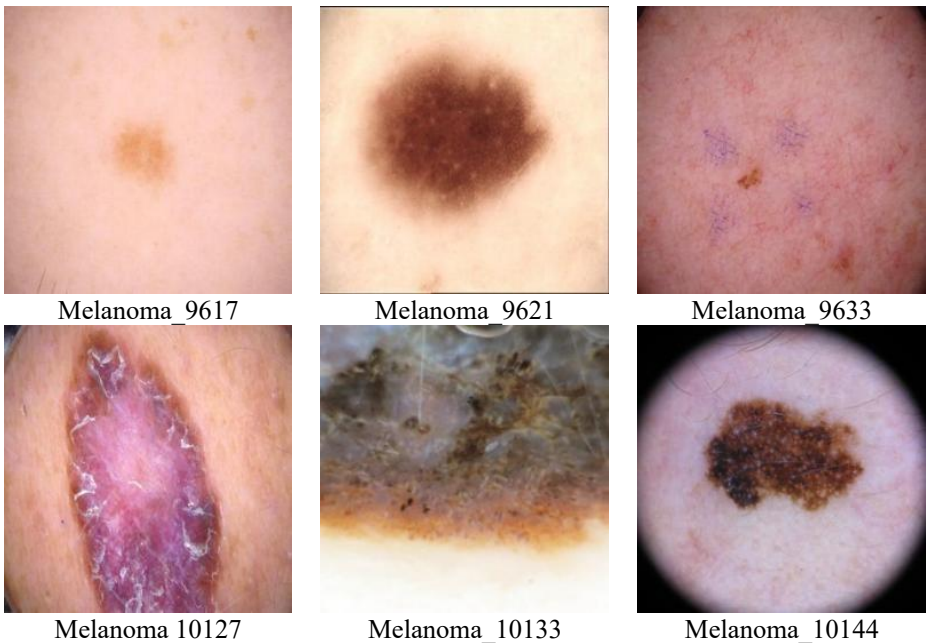


Figure 3. Six images of benign and malignant cases from the test dataset

Table 1. Distribution of dermoscopic image dataset splits for training, validation and testing

Split	Source	Benign	Malignant	Total	Role
Train	(ISIC + HAM1000)/Kaggle	4000	3684	7684	Model training
Validation	(ISIC + HAM1000)/Kaggle	1000	921	1921	Model validation
Test	(ISIC + HAM1000)/Kaggle	500	500	1000	Final evaluation

4.3 Model training and fine-tuning

Each base model is initialized with ImageNet weights and fine-tuned using a custom classification head consisting of global average pooling, batch normalization, and a dense layer with 512 neurons, dropout, and a sigmoid output layer. The Adam optimizer is used because it provides adaptive learning rates and stable gradient updates in non-convex training problems. Binary cross-entropy is employed as the loss function. Early stopping based on validation ROC-AUC is used to limit overfitting, while dropout regularization and batch normalization help improve generalization. These choices are consistent with strategies commonly reported in transfer-learning approaches for medical image classification [12, 17].

A training strategy with a batch size of 32 was employed to optimize memory utilization and improve the convergence behavior of the three deep neural networks.

The combination of frozen backbone layers and selective fine-tuning helps preserve generic visual representations while adapting the final layers to dermoscopic texture and color patterns. Similar studies report that targeted fine-tuning can improve classification performance compared with training strategies that do not adequately adapt pretrained representations to the target domain.

4.4 Stacking ensemble and threshold optimization

To prevent information leakage, training meta-features were generated using a five-fold stratified out-of-fold (OOF) procedure. In each fold, the CNN base learners (MobileNetV2, EfficientNetV2B0, and DenseNet169) were trained on the remaining four folds and used to predict the held-out fold. The resulting OOF predictions were concatenated to form the meta-feature matrix for Logistic Regression and XGBoost. XGBoost was used to capture nonlinear relationships among base-model outputs, while Logistic Regression provided a linear alternative.

The meta-learners were trained exclusively on OOF features, while validation and test predictions were generated separately. The test set was not used for training, hyperparameter optimization, or threshold selection and was reserved for final evaluation. A target sensitivity of 95% was specified as the operating criterion during validation-based threshold selection. Once selected, the threshold was fixed and applied unchanged to the test set, where the resulting sensitivity was evaluated without further threshold adjustment.

Stacking enables the ensemble to learn the relative strengths of the base learners rather than treating them equally, providing a flexible alternative to fixed voting rules [14, 18]. The 95% sensitivity target further prioritizes melanoma detection while maintaining an appropriate balance between sensitivity and specificity.

4.5 Performance evaluation

Model performance is evaluated using accuracy, sensitivity, specificity, precision, F1-score, and ROC-AUC. These metrics are necessary in melanoma detection because a single global measure can conceal clinically important trade-offs. Sensitivity captures the model's ability to identify malignant cases, while specificity indicates how well benign cases are correctly classified. Precision and F1-score provide additional insight into the quality of positive predictions, and ROC-AUC summarizes overall discriminative ability across decision thresholds. The evaluation framework therefore aligns with the diagnostic priorities of medical AI research and enables a more nuanced assessment of ensemble performance [7, 18].

Reliability is further supported by the use of a standardized evaluation pipeline. When a classifier is intended for medical decision support, reporting multiple metrics is essential because high accuracy may coexist with clinically unacceptable false-negative rates. As demonstrated in previous melanoma detection studies, the practical value of melanoma detection models depends on balanced performance and explicit attention to missed malignant lesions [10]. The selected metrics therefore provide a comprehensive summary of diagnostic quality.

4.6 Reproducibility and computational setup

The complete stacking pipeline is saved as a single serialized file so that the preprocessing steps, trained CNN models, feature scaler, and XGBoost meta-learner can be reproduced consistently. Training is conducted on a high-performance workstation with GPU acceleration, which is necessary for fitting multiple deep models efficiently. To make the experiment reproducible, the training parameters, architecture choices, and evaluation metrics are explicitly specified. Such transparency is important in medical machine learning because reproducibility is a prerequisite for reliable model validation and future clinical translation.

Reproducibility also supports scientific comparability. Because melanoma studies often differ in datasets, preprocessing, and thresholding, clear methodological reporting is necessary for meaningful comparison across systems. Reviews of medical image AI repeatedly note that differences in datasets, preprocessing, validation strategies, and experimental protocols can complicate meaningful comparison between models [5, 12, 16]. The present methodology therefore documents not only the core architecture, but also the training logic and evaluation pipeline that make the system usable for future replication and extension.

4.7 Hyperparameter settings of the XGBoost and Logistic Regression meta-learners

Table 2 summarizes the hyperparameter settings used for the XGBoost and Logistic Regression meta-learners. These configurations were selected to provide robust learning, reduce overfitting, and ensure stable convergence during the stacking process.

Table 2. Hyperparameter configuration of the XGBoost and Logistic Regression meta-learners

Type	Hyperparameter	Value
XGBoost	Number of estimators (n_estimators)	500
	Learning rate	0.02
	Maximum tree depth (max_depth)	4
	Subsample ratio	0.8
	Column sample ratio (colsample_bytree)	0.8
	Evaluation metric	ROC-AUC
	Class imbalance weight (scale_pos_weight)	Computed from the training data
	Early stopping rounds	50
	Verbosity	0
Logistic Regression	max_iter	3000
	random_state	24

Note: ROC-AUC = area under the receiver operating characteristic curve.

5. RESULTS AND DISCUSSION

5.1 Performance under different learning rate and weight decay settings

The experimental results obtained under different learning rate and weight decay configurations are summarized in Table 3. The evaluation includes stacking ensembles using Logistic Regression and XGBoost meta-learners combined with both sensitivity-oriented threshold optimization and the Youden index.

The experimental results for Group G1 demonstrate that the proposed stacking ensemble achieved competitive melanoma classification performance under the sensitivity-oriented threshold optimization strategy with Logistic Regression as the meta-learner. Using the optimal hyperparameter setting (learning rate = 2×10^{-4}, weight decay = 1×10^{-5}), the G1 configuration achieved 91.50% accuracy, 91.40% sensitivity, 91.60% specificity, 91.58% precision, 91.49% F1-score, and 97.99% ROC-AUC. The corresponding performance is illustrated in Figure 4, which summarizes the classification metrics obtained by the proposed stacking ensemble. These results demonstrate a good balance between melanoma detection and false-positive reduction, highlighting the effectiveness of the proposed stacking ensemble for reliable melanoma classification.

The XGBoost-based stacking ensemble demonstrated competitive performance across the evaluated hyperparameter configurations. For Group G2, the best 20-epoch configuration, illustrated in Figure 5, achieved 92.20% accuracy, 91.20% sensitivity, 93.20% specificity, 92.12% F1-score, and 97.77% ROC-AUC. G3, obtained with 40 epochs and illustrated in Figure 6, achieved 92.00% accuracy, 92.20% sensitivity, 91.80% specificity, 92.02% F1-score, and 97.74% ROC-AUC. The similar ROC-AUC values indicate comparable discriminative performance across the two training durations. G3 provided slightly higher sensitivity, whereas the difference in ROC-AUC was not statistically significant.

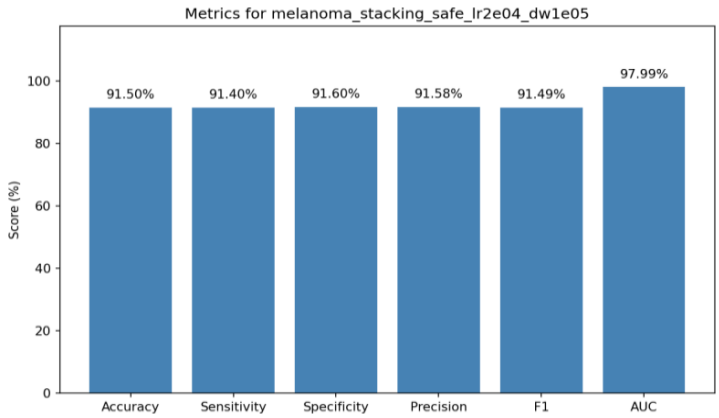


Figure 4. Performance metrics of the Logistic Regression meta-learner (learning rate = 2×10^{-4}; weight decay = 1×10^{-5})

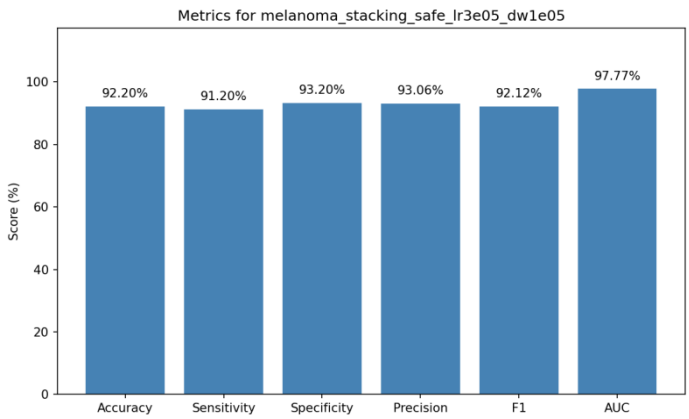


Figure 5. Performance metrics of the XGBoost meta-learner (learning rate = 3×10^{-5}; weight decay = 1×10^{-5}; epochs = 20)

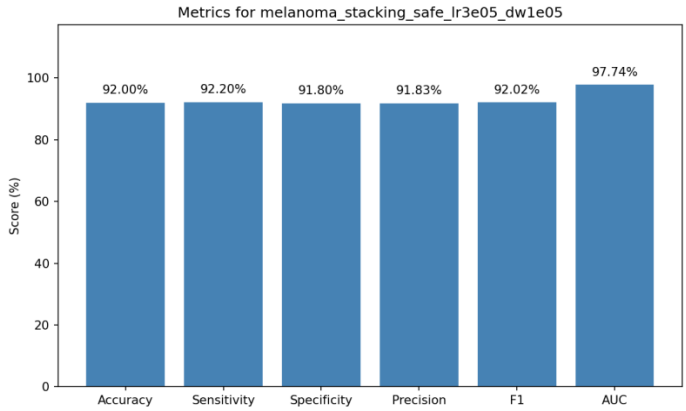


Figure 6. Performance metrics of the XGBoost meta-learner (learning rate = 3×10^{-5}; weight decay = 1×10^{-5}; epochs = 40)

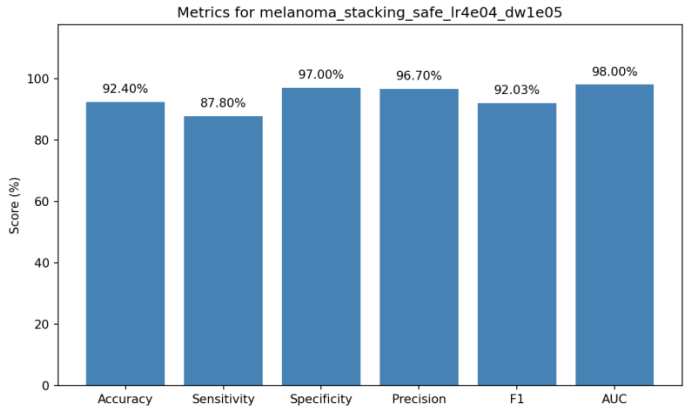


Figure 7. Performance metrics of the Logistic Regression meta-learner (learning rate = 4×10^{-4}; weight decay = 1×10^{-5})

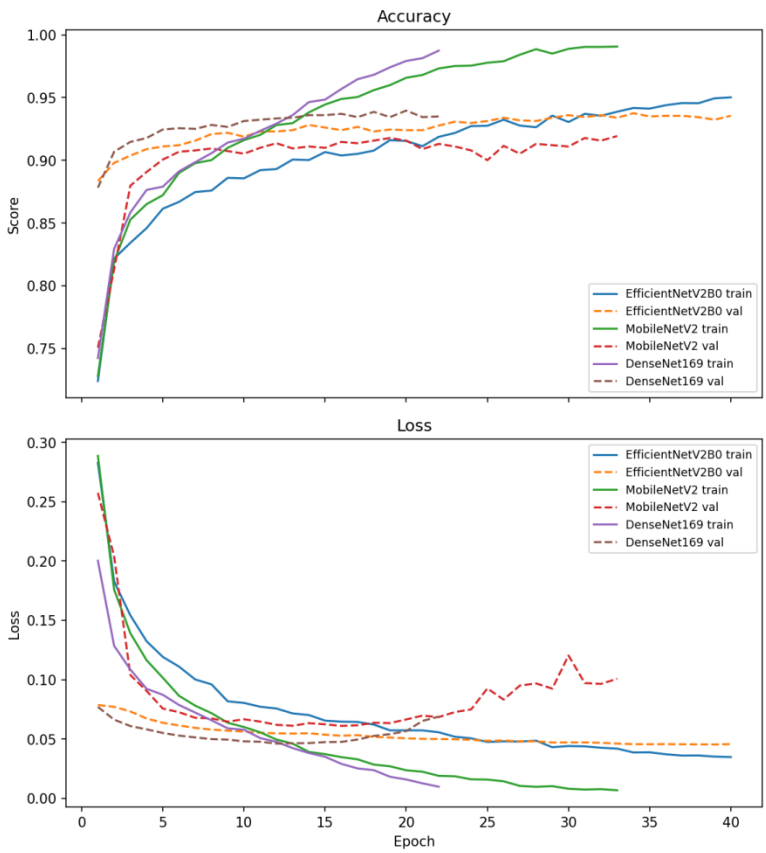


Figure 8. Training and validation (loss and accuracy) of EfficientNetV2B0, MobileNetV2, and DenseNet169 for Group G3

Table 3. Performance of the stacking ensemble under different learning rate and weight decay configurations

Group	Epochs	Learning Rate	Weight Decay	Model	Accuracy	Sensitivity	Specificity	Precision	F1-Score	ROC-AUC
G1	20	2×10^{-4}	1×10^{-5}	Meta learner	91.50%	91.40%	91.60%	91.58%	91.49%	97.99%
				EfficientNetV2B0	91.00%	89.00%	93.00%	92.71%	90.82%	97.97%
				MobileNetV2	88.50%	92.40%	84.60%	85.71%	88.93%	96.68%
				DenseNet169	91.40%	88.20%	94.60%	94.23%	91.12%	97.43%
G2	20	3×10^{-5}	1×10^{-5}	Meta learner	92.20%	91.20%	93.20%	93.06%	92.12%	97.77%
				EfficientNetV2B0	90.70%	87.60%	93.80%	93.39%	90.40%	97.43%
				MobileNetV2	90.20%	91.40%	89.00%	89.26%	90.32%	96.95%
				DenseNet169	91.70%	88.00%	95.40%	95.03%	91.38%	97.38%
G3	40	3×10^{-5}	1×10^{-5}	Meta learner	92.00%	92.20%	91.80%	91.83%	92.02%	97.74%
				EfficientNetV2B0	91.80%	88.40%	95.20%	94.85%	91.51%	97.40%
				MobileNetV2	91.60%	90.60%	92.60%	92.45%	91.52%	97.40%
				DenseNet169	91.40%	87.80%	95.00%	94.61%	91.08%	97.38%
G4	20	4×10^{-4}	1×10^{-5}	Meta learner	92.40%	87.80%	97.00%	96.70%	92.03%	98.00%
				EfficientNetV2B0	92.10%	90.40%	93.80%	93.58%	91.96%	97.92%
				MobileNetV2	90.30%	85.40%	95.20%	94.68%	89.80%	96.66%
				DenseNet169	91.10%	91.40%	90.80%	90.85%	91.13%	97.12%

Note: G1: Sensitivity-oriented threshold optimization and Meta-Linear = Logistic Regression (20 epochs); G2: Sensitivity-oriented threshold optimization and Meta-Linear = XGBoost (20 epochs); G3: Sensitivity-oriented threshold optimization and Meta-Linear = XGBoost (40 epochs); G4: Youden index-based threshold optimization and Meta-Linear = Logistic Regression. ROC-AUC = area under the receiver operating characteristic curve (20 epochs).

Table 4. Performance comparison between individual convolutional neural network (CNN) models and the proposed stacking (XGBoost meta-learner)

Model	Accuracy	Sensitivity	Specificity	Precision	F1-Score	ROC-AUC
EfficientNetV2B0	91.80%	88.40%	95.20%	94.85%	91.51%	97.40%
MobileNetV2	91.60%	90.60%	92.60%	92.45%	91.52%	97.40%
DenseNet169	91.40%	87.80%	95.00%	94.61%	91.08%	97.38%
XGBoost Meta-Learner	92.00%	92.20%	91.80%	91.83%	92.02%	97.74%

Note: ROC-AUC = area under the receiver operating characteristic curve.

Under the Youden index criterion, Group G4 (learning rate = 4×10^{-4}, weight decay = 1×10^{-5}), presented in Figure 7, achieved 92.40% accuracy, 87.80% sensitivity, 97.00% specificity, and 98.00% ROC-AUC. Compared with G3 (92.00% accuracy, 92.20% sensitivity, 91.80% specificity, and 97.74% ROC-AUC), G4 improved specificity and ROC-AUC but reduced sensitivity by 4.40 percentage points. Given the importance of sensitivity in melanoma detection, G3 was preferred as the final configuration because it achieved higher sensitivity (92.20%) than G4 (87.80%). In contrast, G4 achieved higher specificity and a slightly higher ROC-AUC. The ROC-AUC difference between the two configurations was not statistically significant.

Overall, the experimental results show that the evaluated stacking configurations achieved competitive performance. G3 achieved the highest sensitivity among the four groups (92.20%), together with 92.00% accuracy, 91.80% specificity, 92.02% F1-score, and 97.74% ROC-AUC. G4 achieved higher specificity (97.00%) and a slightly higher ROC-AUC (98.00%) but lower sensitivity (87.80%). Therefore, G3 was selected as the final proposed model because its higher sensitivity better matches the primary objective of melanoma detection. The DeLong analysis indicated that the differences in ROC-AUC among the evaluated pipelines were not statistically significant.

Figure 8 shows smooth convergence for the three CNN models. EfficientNetV2B0 and DenseNet169 achieved higher validation accuracy than MobileNetV2, while DenseNet169 exhibited stable convergence. Both training and validation losses decreased consistently, indicating effective optimization with limited overfitting. The 40-epoch configuration also yielded slightly higher sensitivity than the 20-epoch setting, while the ROC-AUC values remained comparable.

5.2 Interpretation of ensemble benefits

To assess the contribution of the proposed stacking framework, the XGBoost meta-learner was systematically compared with the individual CNN base models under the same training configuration as G3 in Table 3 (40 epochs, learning rate = 3×10^{-5}, and weight decay = 1×10^{-5}). The comparative results are summarized in Table 4.

The proposed stacking XGBoost meta-learner G3 achieved an accuracy of 92.00%, sensitivity of 92.20%, F1-score of 92.02%, and ROC-AUC of 97.74%. Compared with EfficientNetV2B0, the proposed stacking model showed higher sensitivity by 3.80 percentage points, F1-score by 0.51 percentage points, and ROC-AUC by 0.34 percentage points. Relative to MobileNetV2, improvements of 0.40% in accuracy, 1.60% in sensitivity, 0.50% in F1-score, and 0.34% in ROC-AUC were obtained. likewise, compared with DenseNet169, the stacking model increased accuracy by 0.60%, sensitivity by 4.40%, F1-score by 0.94%, and ROC-AUC by 0.36%.

Although EfficientNetV2B0 and DenseNet169 achieved slightly higher specificity and precision, these gains were accompanied by lower sensitivity, leading to a greater number of missed melanoma cases. In contrast, the proposed stacking framework provided a better balance between sensitivity and specificity, resulting in the highest F1-score among the evaluated configurations, while the ROC-AUC differences were not statistically significant. This behavior indicates that combining the complementary feature representations learned by different CNN architectures enables the meta-learner to compensate for the weaknesses of individual models while

preserving their strengths. Consequently, the ensemble produces more robust and reliable predictions than any single CNN, making it particularly suitable for melanoma diagnosis, where maximizing lesion detection while maintaining good overall accuracy is of primary clinical importance.

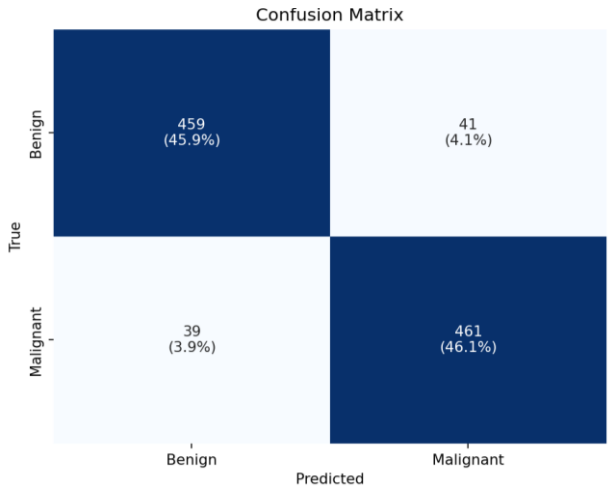


Figure 9. Confusion matrix of the best XGBoost stacking model (40 epochs)

Figure 9 also demonstrates that the proposed ensemble correctly classified the vast majority of melanoma and benign lesions. Only a limited number of false negatives were observed, confirming the high sensitivity of the proposed framework for clinical melanoma detection.

The results also demonstrate that ROC-based threshold optimization using the sensitivity-oriented threshold optimization criterion provides a more clinically meaningful trade-off than arbitrary default cutoffs. Prioritizing high sensitivity is particularly important in melanoma diagnosis because missing malignant lesions may have severe clinical consequences. The optimized threshold therefore improves the practical diagnostic utility of the framework by maintaining strong melanoma detection capability while preserving acceptable specificity.

To provide a detailed assessment of the classification performance, the confusion matrix of the best-performing stacking model on the test set is presented in Table 5. The matrix reports the numbers of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN), thereby complementing the summary performance metrics.

As shown in Table 5, the proposed model correctly classified 459 benign and 461 malignant lesions. Only 41 benign lesions were incorrectly classified as malignant (false positives), while 39 malignant lesions were misclassified as benign (false negatives), demonstrating a balanced trade-off between sensitivity and specificity.

Table 5. Confusion matrix of the selected final model on the test set (n = 1,000)

Actual Class	Predicted Benign	Predicted Malignant
Benign (n = 500)	459 (TN)	41 (FP)
Malignant (n = 500)	39 (FN)	461 (TP)

5.3 Comparison of all meta-learner groups

Four experimental groups were evaluated using different optimization strategies and meta-learners, as shown in Table 3: (G1) sensitivity-oriented threshold optimization with Logistic Regression, (G2) sensitivity-oriented threshold optimization with XGBoost (20 epochs), (G3) sensitivity-oriented threshold optimization with XGBoost (40 epochs), and (G4) Youden index-based threshold optimization with Logistic Regression. For G4, the Youden index was used to identify the operating threshold that provides a balance between sensitivity and specificity [19]. Since the primary objective of this study is melanoma detection, the comparison emphasizes sensitivity, followed by ROC-AUC, F1-score, specificity, and accuracy, given the importance of minimizing false-negative cases.

Figure 10 illustrates the ROC curves of the best-performing models from the four experimental groups. All evaluated configurations showed high numerical discriminative performance, with ROC-AUC values above 97%. The Youden index-based Logistic Regression model (G4) achieved the highest numerical ROC-AUC (98.00%), whereas the XGBoost-based stacking model with 40 epochs (G3) achieved an ROC-AUC of 97.74% and the highest sensitivity (92.20%). Since the DeLong test showed that the differences in ROC-AUC among the evaluated pipelines were not statistically significant [23], G3 was selected based on its higher sensitivity at the selected operating point.

As shown in Figure 11, G1 achieved an ROC-AUC of 97.99%, with 91.50% accuracy, 91.40% sensitivity, and 91.49% F1-score. These results indicate that Logistic Regression provided competitive performance as a linear meta-learner. However, the XGBoost configurations achieved higher sensitivity or accuracy at the selected operating points.

The 20-epoch XGBoost configuration achieved the highest accuracy (92.20%) among all sensitivity-constrained models while maintaining a high sensitivity (91.20%) and F1-score (92.12%). This performance can be attributed to XGBoost's ability to model nonlinear relationships among the predictions of the base learners, providing an alternative to the linear decision mechanism of logistic regression. Increasing the training duration to 40 epochs resulted in G3, which achieved a higher sensitivity of 92.20% compared with G2 (91.20%), while maintaining a comparable ROC-AUC of 97.74%. The corresponding XGBoost meta-learner (G3) achieved the highest sensitivity (92.20%), while maintaining an Accuracy of 92.00%, an F1-score of 92.02%, and an ROC-AUC of 97.74%. Although its accuracy is slightly lower than that of the 20-epoch configuration, the higher sensitivity indicates that more melanoma cases are correctly identified, which is particularly relevant to the study objective of minimizing false-negative melanoma cases. The Youden index + Logistic Regression group (G4) obtained the highest accuracy (92.40%), specificity (97.0%), precision (96.70%), and ROC-AUC (98.00%), showing the highest numerical ROC-AUC among the evaluated configurations, although the difference was not statistically significant. However, its sensitivity decreased to 87.8%, implying that approximately one melanoma out of eight could be missed. While such a trade-off may be acceptable for balanced binary classification tasks, it is less desirable for melanoma screening, where false negatives have potentially life-threatening consequences.

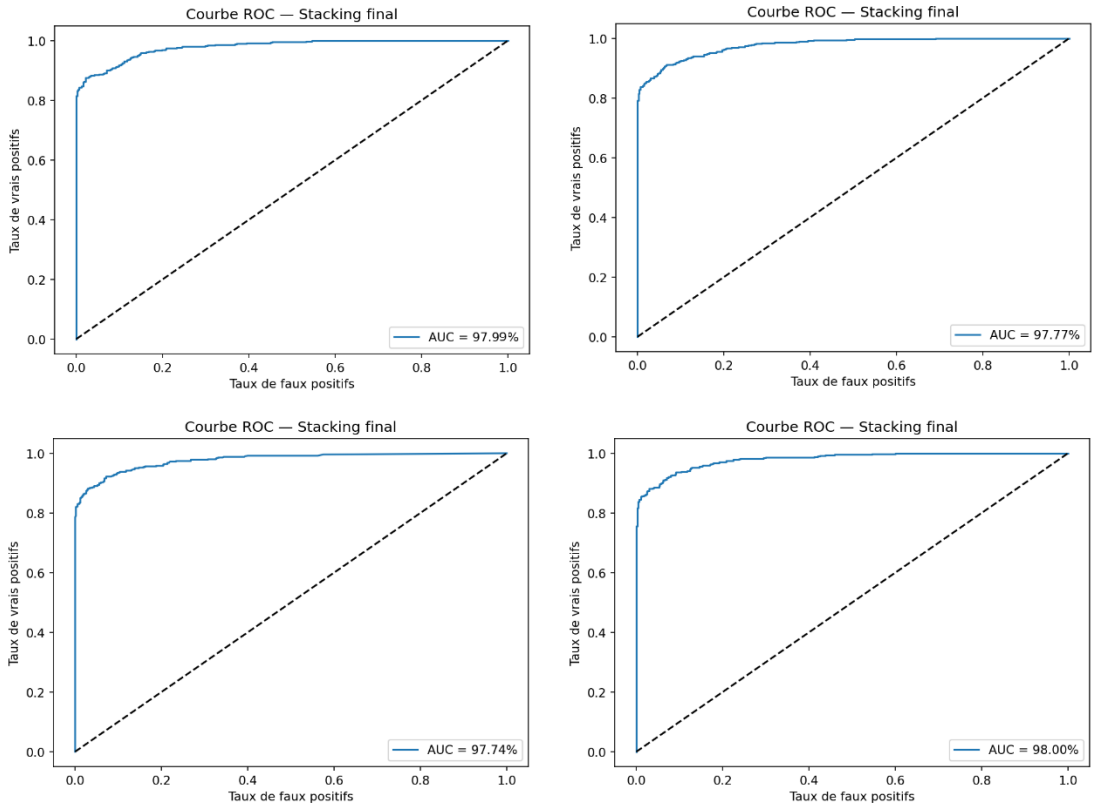


Figure 10. Receiver operating characteristic (ROC) curves of the four best meta-learner models

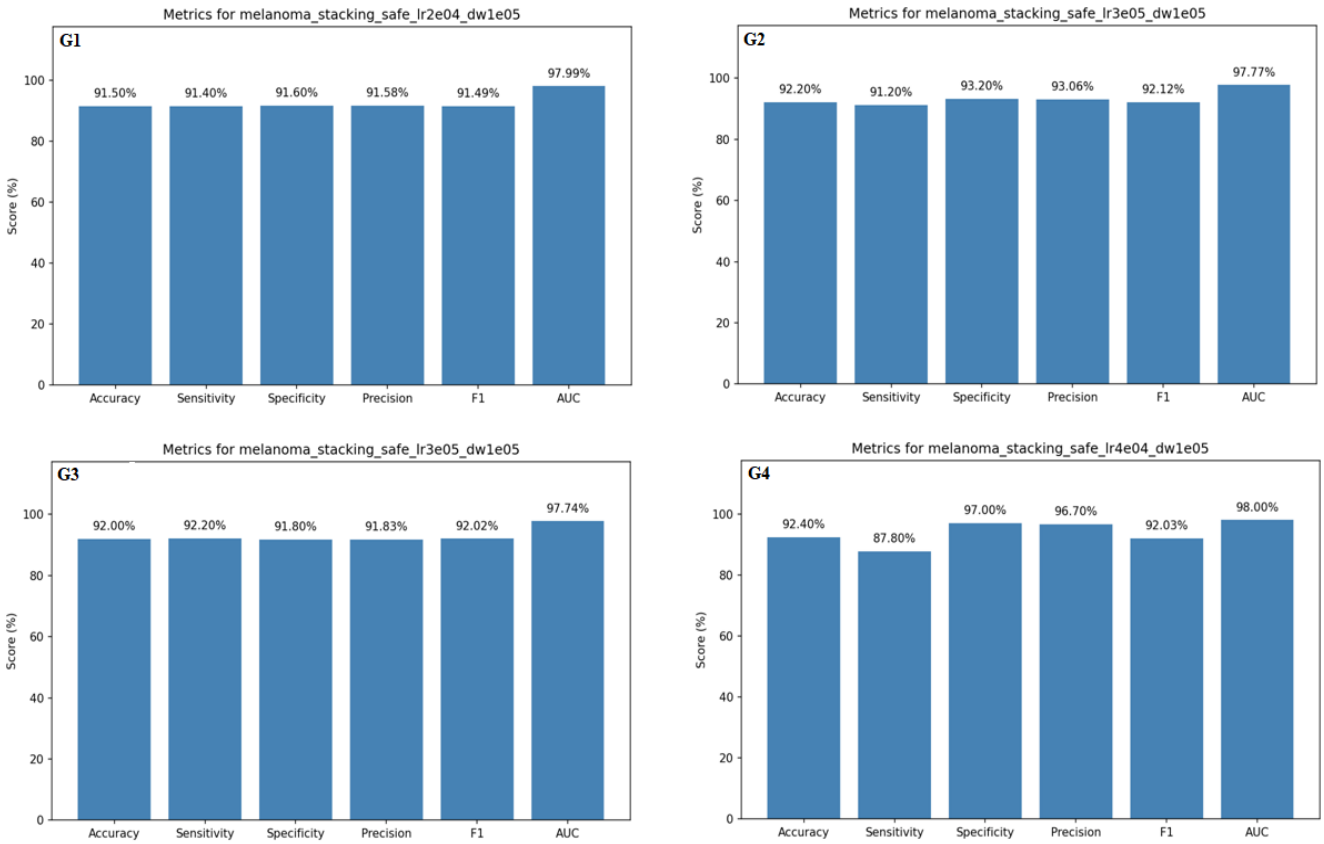


Figure 11. Overall comparison of the four meta-learner groups

Overall, the comparison indicates that the sensitivity-oriented strategy provides a higher sensitivity at the selected operating point than the Youden index strategy. This characteristic is particularly relevant when the primary objective is to minimize false-negative melanoma cases. The sensitivity-oriented optimization resulted in a higher sensitivity at the selected operating point, while the observed ROC-AUC differences were not statistically significant.

5.4 Model calibration analysis

Table 6 reports the Brier Scores obtained by the evaluated models. Among the individual CNNs, MobileNetV2 achieved the lowest Brier Score (0.0617), indicating the most accurate probability estimates. DenseNet169 and EfficientNetV2B0 also demonstrated good calibration with Brier Scores below 0.07.

Table 6. Brier scores of the evaluated models

Model	Brier Score
EfficientNetV2B0	0.0673
MobileNetV2	0.0617
DenseNet169	0.0655
Proposed Stacking Model	0.1048

Figure 12 illustrates the corresponding calibration curves. The individual CNN models generally follow the ideal calibration line, whereas the stacking classifier exhibits a larger deviation for intermediate probability values. Although the proposed stacking model achieved a high ROC-AUC of 97.74%, its Brier Score (0.1048) was higher than those of the individual CNN models, indicating less accurate probability calibration. These results suggest that the stacking configuration provided strong discriminative performance but did not improve probability calibration relative to the individual CNNs.

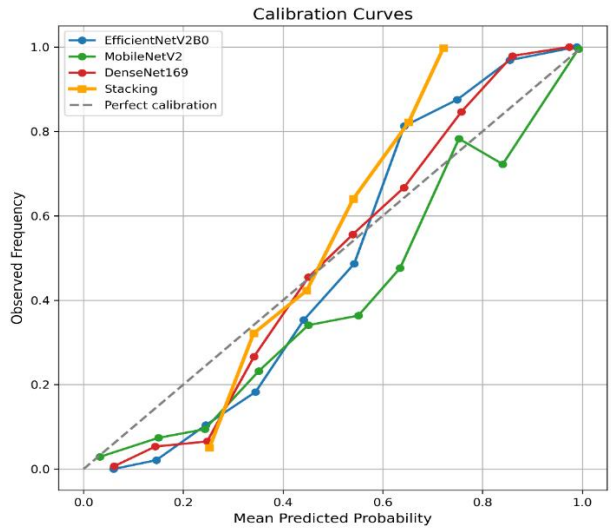


Figure 12. Calibration curves of the evaluated models

Table 1. Pairwise DeLong test results for ROC-AUC comparison of the proposed stacking pipelines

Comparaison	ROC-AUC	*p*-Value	Conclusion
Pipeline 1 vs Pipeline 2	97.99% vs 97.77%	0.0605	Not significant
Pipeline 1 vs Pipeline 3	97.99% vs 97.74%	0.1098	Not significant
Pipeline 1 vs Pipeline 4	97.99% vs 98.00%	0.9140	Not significant
Pipeline 2 vs Pipeline 3	97.77% vs 97.74%	0.8137	Not significant
Pipeline 2 vs Pipeline 4	97.77% vs 98.00%	0.1857	Not significant
Pipeline 3 vs Pipeline 4	97.74% vs 98.00%	0.1584	Not significant

Note: Pipeline 1 = G1, Pipeline 2 = G2, Pipeline 3 = G3, Pipeline 4 = G4. ROC-AUC = area under the receiver operating characteristic curve.

5.5 DeLong test for ROC-AUC comparison

To statistically assess the discriminative performance of the proposed stacking pipelines, pairwise ROC-AUC comparisons were performed using DeLong's test. This nonparametric test evaluates whether differences between correlated ROC curves are statistically significant. The results of all pairwise comparisons among the four pipelines are summarized in Table 7.

As shown in Table 7, all pairwise comparisons yielded p-values greater than the predefined significance level of 0.05, indicating that no statistically significant differences in ROC-AUC were observed among the four evaluated pipelines. The largest numerical difference in ROC-AUC was observed between Pipeline 3 (G3; ROC-AUC = 97.74%) and Pipeline 4 (G4; ROC-AUC = 98.00%). However, this difference was not statistically significant (p = 0.1584). Similarly, the comparisons between Pipeline 1 and Pipeline 2 (p = 0.0605), Pipeline 1 and Pipeline 3 (p = 0.1098), Pipeline 1 and Pipeline 4 (p = 0.9140), Pipeline 2 and Pipeline 3 (p = 0.8137), and Pipeline 2 and Pipeline 4 (p = 0.1857) all yielded p-values above 0.05.

Overall, these results indicate that the four evaluated stacking configurations exhibited comparable ROC-AUC performance on the test dataset. Therefore, the selection of G3 as the final proposed model was not based on statistically significant superiority in ROC-AUC, but rather on its higher sensitivity at the selected operating point, which is more consistent with minimizing false-negative melanoma classifications.

5.6 Statistical validation using bootstrap confidence intervals

To evaluate the robustness of the proposed stacking pipelines, a bootstrap analysis was conducted to estimate the 95% confidence intervals (CIs) of the ROC-AUC [24]. The results of the three base CNN models and the proposed stacking model are presented in Table 8.

Table 8. Area under the ROC curve with 95% bootstrap confidence intervals (CIs)

Model	ROC-AUC (%)	95% Bootstrap CI
EfficientNetV2B0	97.40	96.55–98.13
MobileNetV2	97.40	96.56–98.16
DenseNet169	97.38	96.49–98.16
Proposed Stacking Model	97.74	96.90–98.45

Note: ROC-AUC = area under the receiver operating characteristic curve.

To assess the statistical reliability of the reported ROC-AUC values, 95% CIs were estimated using bootstrap resampling with 2,000 iterations. As shown in Table 8, the proposed stacking model achieved the highest ROC-AUC of 97.74% (95% CI: 96.90–98.45%), whereas the individual CNN models achieved ROC-AUC values ranging from

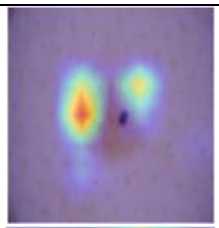
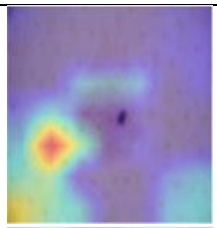
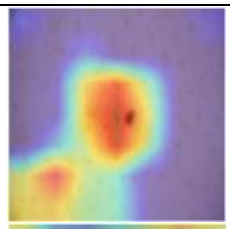
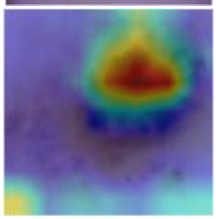
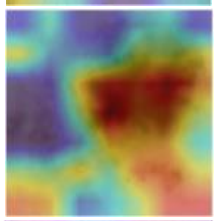
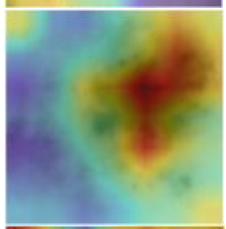
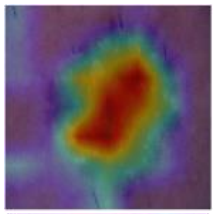
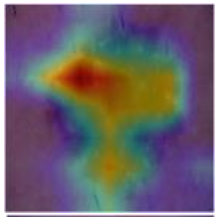
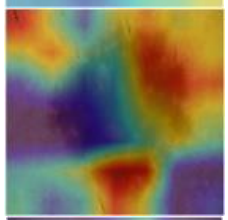
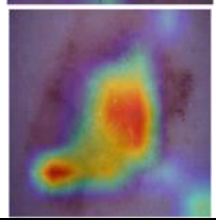
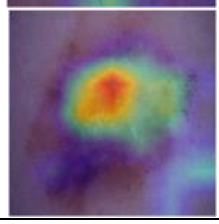
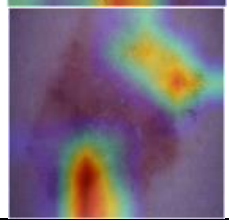
97.38% to 97.40%. The relatively narrow CIs indicate stable discrimination performance across bootstrap samples.

5.7 Explainability analysis using Grad-CAM

To improve the interpretability of the proposed stacking framework, Grad-CAM [25] was employed to visualize the image regions that contributed most to the predictions of the three CNN backbones (EfficientNetV2B0, MobileNetV2, and DenseNet169). The generated activation maps provide qualitative evidence that the models focus primarily on clinically relevant lesion areas rather than surrounding background regions.

Table 9 presents Grad-CAM visualizations for four representative dermoscopic images, including both benign and malignant lesions, together with the corresponding final stacking predictions. The visualizations were generated independently for the three CNN base learners, namely EfficientNetV2B0, MobileNetV2, and DenseNet169. Although the activation patterns vary across architectures, the three CNNs generally focus on lesion regions and visually salient structures, including irregular borders, asymmetric regions, and highly pigmented areas. These observations provide qualitative evidence that the individual CNNs learn discriminative representations from relevant lesion regions.

Table 9. Grad-CAM visualizations of the three convolutional neural network (CNN) base learners and corresponding final stacking predictions for representative dermoscopic images

Original	EfficientNetV2B0	MobileNetV2	DenseNet169	XGBoost Stacking Prediction
				Stack: 0.267 True: benign Prediction: benign
				Stack: 0.722 True: malignant Prediction: malignant
				Stack: 0.648 True: benign Prediction: malignant
				Stack: 0.325 True: malignant Prediction: benign

The quantitative results in Table 4 further characterize the performance of the individual CNN base learners. EfficientNetV2B0, MobileNetV2, and DenseNet169 achieved comparable ROC-AUC values of 97.40%, 97.40%, and 97.38%, respectively, with accuracies ranging from 91.40% to 91.80%. EfficientNetV2B0 and DenseNet169 achieved higher specificity values of 95.20% and 95.00%, respectively, whereas MobileNetV2 achieved the highest sensitivity among the individual CNNs (90.60%). These results indicate differences in the predictive characteristics captured by the three CNN architectures.

The predictions generated by the three CNN base learners were subsequently combined by the XGBoost meta-learner in the proposed stacking framework. The final G3 configuration achieved 92.00% accuracy, 92.20% sensitivity, 92.02% F1-score, and 97.74% ROC-AUC. These results indicate competitive performance of the stacking configuration relative to the individual CNN models. Importantly, the Grad-CAM visualizations represent the CNN base learners only and should not be interpreted as a direct visualization of the XGBoost meta-learner. They provide qualitative insight into the image regions considered by the individual CNNs before their predictions are combined by the stacking model.

The final stacking prediction is reported together with its predicted probability and corresponding classification outcome (TP, TN, FP, or FN). Overall, the Grad-CAM analysis complements the quantitative evaluation by providing qualitative interpretability of the CNN base learners and illustrating differences in their activation patterns. However, these visualizations do not directly explain the decision mechanism of the XGBoost meta-learner. The final stacking decision is therefore assessed through the quantitative performance of the complete ensemble.

5.8 Selection of the optimal model

Considering the primary objective of melanoma detection, the XGBoost-based stacking model with 40 epochs (G3) was selected as the final proposed model. This configuration uses MobileNetV2, EfficientNetV2B0, and DenseNet169 as base learners and XGBoost as the meta-classifier.

This selection is justified by the following observations:

(1) It achieved the highest sensitivity among the evaluated configurations (92.20%), which is particularly relevant for melanoma detection because of the importance of minimizing false-negative cases.

(2) It maintained a high accuracy (92.00%), indicating excellent overall classification performance.

(3) Its F1-score (92.02%) demonstrates a strong balance between precision and sensitivity.

(4) The ROC-AUC (97.74%) confirms strong discriminative performance.

(5) Although the Logistic Regression model optimized using the Youden index (G4) achieved a slightly higher ROC-AUC of 98.00% and specificity of 97.00%, its sensitivity was lower at 87.80%.

Therefore, G3 was selected because its higher sensitivity better matches the primary objective of melanoma detection. The DeLong test indicated that the differences in ROC-AUC among the evaluated pipelines were not statistically significant.

6. CONCLUSIONS

This study presented a stacking-based deep learning framework for melanoma detection using MobileNetV2, EfficientNetV2B0, and DenseNet169 as base learners and XGBoost as the meta-classifier. The experimental results show that the proposed ensemble achieved competitive performance across the evaluated configurations. In particular, the G3 configuration provided a favorable operating point, achieving 92.20% sensitivity, 92.02% F1-score, and 97.74% ROC-AUC. These results indicate that the stacking framework can effectively combine complementary predictions from different CNN architectures. However, the DeLong analysis showed that the differences in ROC-AUC among the compared pipelines were not statistically significant. Therefore, the observed performance differences should be interpreted as improvements at the selected operating point rather than evidence of statistically significant superiority in overall discriminative ability.

More broadly, the study demonstrates the feasibility of integrating lightweight CNN architectures into a stacking framework while maintaining competitive classification performance. The ROC-based threshold optimization provides flexibility in selecting an operating point according to the intended diagnostic objective, particularly when sensitivity is prioritized to reduce false-negative predictions. The results obtained on the test dataset provide evidence supporting the potential of the proposed methodology, while recognizing that further validation is required.

The findings also highlight the value of combining complementary predictions from different CNN architectures through a nonlinear meta-learner. The selected G3 configuration achieved a favorable balance between sensitivity and other classification metrics, suggesting that the sensitivity-oriented operating point may be appropriate when the primary objective is to improve melanoma detection. Nevertheless, the absence of statistically significant ROC-AUC differences indicates that the proposed ensemble should not be considered statistically superior to the compared pipelines in terms of overall discriminative ability.

Despite its strong performance, the proposed framework has several limitations that should be acknowledged. First, the model depends heavily on the quality and diversity of the training data, which may limit generalization when applied to external clinical populations with different acquisition conditions or demographic distributions. Second, although the stacking strategy improves robustness, it also increases computational complexity relative to a single-model baseline, particularly during training. Third, the approach remains largely image-based and does not incorporate clinical metadata such as patient age, lesion history, or anatomical location, which may contain valuable diagnostic information. Fourth, interpretability remains limited. While XGBoost is more transparent than many meta-learning alternatives, the underlying CNN feature representations remain difficult to explain in clinically intuitive terms, which may reduce trust among practitioners if the system is deployed without accompanying explanation methods such as saliency mapping or attention visualization. Finally, the current evaluation does not include prospective clinical testing, meaning that further validation is needed before practical adoption. Future work should therefore focus on external validation, explainable AI methods, and multimodal integration to strengthen real-world utility.

ACKNOWLEDGMENTS

The authors thank the ICOSI Laboratory, Abbes Laghrour University, for providing the computational resources. Figures 1 and 2 were generated using ChatGPT (OpenAI) and DALL·E, then reviewed and refined by the authors, who take full responsibility for the final content.

DATA AND CODE AVAILABILITY

The source code and implementation details of the proposed melanoma detection framework are publicly available on the GitHub Repository: https://github.com/akachatbrahim/Stacking_CNN_Melanoma.git.

The trained deep learning models and stacking ensemble models used in this study are publicly available through Kaggle Models: https://www.kaggle.com/api/v1/datasets/download/brahimakachat/models-melanoma-detection.

The melanoma dataset used in this study is publicly available through the Melanoma Skin Cancer Dataset of 10,605 Images on Kaggle. The dataset was compiled by Muhammad Hasnain Javid from publicly available skin lesion repositories, primarily the ISIC Archive and the HAM10000 dataset [22]. It was created to provide a large-scale binary classification benchmark for automated melanoma detection using deep learning. Dataset: https://www.kaggle.com/datasets/hasnainjaved/melanoma-skin-cancer-dataset-of-10000-images.

ETHICS STATEMENT

This study exclusively used publicly available dermoscopic images obtained from the Kaggle Melanoma Skin Cancer Dataset, which is derived from the ISIC Archive and the HAM10000 dataset. No human participants were recruited, and no personally identifiable patient information was accessed. The study involved a secondary analysis of existing public datasets only. Therefore, institutional ethical approval and informed consent were not required.

REFERENCES

[1] Esteva, A., Kuprel, B., Novoa, R.A., et al. (2017). Dermatologist-level classification of skin cancer with deep neural networks. Nature, 542(7639): 115-118. https://doi.org/10.1038/nature21056

[2] Haenssle, H.A., Fink, C., Schneiderbauer, R., et al. (2018). Man against machine: Diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists. Annals of Oncology, 29(8): 1836-1842. https://doi.org/10.1093/annonc/mdy166

[3] Codella, N.C.F., Gutman, D., Celebi, M.E., et al. (2018). Skin lesion analysis toward melanoma detection: A challenge at the 2017 International Symposium on Biomedical Imaging (ISBI), hosted by the International Skin Imaging Collaboration (ISIC). In 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018), Washington, DC, USA, pp. 168-172. https://doi.org/10.1109/ISBI.2018.8363547

[4] Papachristou, P., Söderholm, M., Pallon, J., et al. (2024). Evaluation of an artificial intelligence-based decision support for detection of cutaneous melanoma in primary care: A prospective, real-life, clinical trial. British Journal of Dermatology, 191(1): 125-133. https://doi.org/10.1093/bjd/ljae021

[5] Naseri, H., Safaei, A. (2025). Diagnosis and prognosis of melanoma from dermoscopy images using machine learning and deep learning: A systematic literature review. BMC Cancer, 25: 75. https://doi.org/10.1186/s12885-024-13423-y

[6] Sarıateş, M., Özbay, E. (2025). Transfer learning-based ensemble of CNNs and vision transformers for accurate melanoma diagnosis and image retrieval. Diagnostics, 15(15): 1928. https://doi.org/10.3390/diagnostics15151928

[7] Gadag, S., Malode Vishwanathac, P.R., Dalal, V.B. (2025). Melanoma classification using ensemble deep transfer learning. International Journal of Artificial Intelligence, 14(6): 4943-4956. https://doi.org/10.11591/ijai.v14.i6.pp4943-4956

[8] Al Mahmud, A., Azam, S., Khan, I.U., et al. (2024). SkinNet-14: A deep learning framework for accurate skin cancer classification using low-resolution dermoscopy images with optimized training time. Neural Computing and Applications, 36: 18935-18959. https://doi.org/10.1007/s00521-024-10225-y

[9] Navaneethan, S., Shilpa, K., Priyadharshani, S. (2025). Enhancing skin cancer detection: Lightweight ensemble model for real-time dermoscopy image classification. In 2025 International Conference on Next Generation Communication & Information Processing (INCIP), Bangalore, India, pp. 850-855. https://doi.org/10.1109/INCIP64058.2025.11018981

[10] Di Biasi, L., De Marco, F., Citarella, A.A., Castrillón-Santana, M., Barra, P., Tortora, G. (2023). Refactoring and performance analysis of the main CNN architectures: Using false negative rate minimization to solve the clinical images melanoma detection problem. BMC Bioinformatics, 24: 386. https://doi.org/10.1186/s12859-023-05516-5

[11] Møllersen, K., Kirchesch, H., Zortea, M., Schopf, T.R., Hindberg, K., Godtliebsen, F. (2015). Computer-aided decision support for melanoma detection applied on melanocytic and nonmelanocytic skin lesions: A comparison of two systems based on automatic analysis of dermoscopic images. BioMed Research International, 2015: 579282. https://doi.org/10.1155/2015/579282

[12] Baig, R., Bibi, M., Hamid, A., Kausar, S., Khalid, S. (2020). Deep learning approaches towards skin lesion segmentation and classification from dermoscopic images—A review. Current Medical Imaging, 16(5): 513-533. https://doi.org/10.2174/1573405615666190129120449

[13] Young, A.T., Vora, N.B., Cortez, J., et al. (2021). The role of technology in melanoma screening and diagnosis. Pigment Cell & Melanoma Research, 34(2): 288-300. https://doi.org/10.1111/pcmr.12907

[14] Wolpert, D.H. (1992). Stacked generalization. Neural Networks, 5(2): 241-259. https://doi.org/10.1016/S0893-6080(05)80023-1

[15] Chen, T.Q., Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In KDD '16: The 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, California, USA, pp. 785-794. https://doi.org/10.1145/2939672.2939785

[16] Goyal, M., Knackstedt, T., Yan, S.F., Hassanpour, S. (2020). Artificial intelligence-based image classification for diagnosis of skin cancer: Challenges and opportunities. Computers in Biology and Medicine, 127: 104065. https://doi.org/10.1016/j.compbiomed.2020.104065

[17] Mohammed, A.S., Ali, M.S.M., Abdalwahid, S.M.J., Kareem, S.W. (2025). Contrastive self-supervised ensemble transfer learning for robust skin cancer classification and early detection. Asian Pacific Journal of Cancer Prevention, 26(7): 2607-2617. https://doi.org/10.31557/APJCP.2025.26.7.2607

[18] Dorathi Jayaseeli, J.D., Briskilal, J., Fancy, C., et al. (2025). An intelligent framework for skin cancer detection and classification using fusion of Squeeze-Excitation-DenseNet with metaheuristic-driven ensemble deep learning models. Scientific Reports, 15: 7425. https://doi.org/10.1038/s41598-025-92293-1

[19] Youden, W.J. (1950). Index for rating diagnostic tests. Cancer, 3(1): 32-35. https://doi.org/10.1002/1097-0142(1950)3:1<32::AID-CNCR2820030106>3.0.CO;2-3

[20] Alotaibi, A., AlSaeed, D. (2025). Skin cancer detection using transfer learning and deep attention mechanisms. Diagnostics, 15(1): 99. https://doi.org/10.3390/diagnostics15010099

[21] Javid, M.H. (2021). Melanoma skin cancer dataset of 10000 images. Kaggle. https://www.kaggle.com/datasets/hasnainjaved/melano

ma-skin-cancer-dataset-of-10000-images.
[22] Tschandl, P., Rosendahl, C., Kittler, H. (2018). The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. Scientific Data, 5: 180161. https://doi.org/10.1038/sdata.2018.161
[23] DeLong, E.R., DeLong, D.M., Clarke-Pearson, D.L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. Biometrics, 44(3): 837-845. https://doi.org/10.2307/2531595
[24] Efron, B., Tibshirani, R.J. (1993). An Introduction to the Bootstrap. New York, NY: Chapman and Hall.
[25] Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization. In 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, pp. 618-626. https://doi.org/10.1109/ICCV.2017.74