Vol. 16 (2026): Proceedings of the 2026 International Conference on AI and Technological Innovation Applications (AITIA 2026)
E-ISSN: 3105-0522 P-ISSN: 3105-0514
DOI: https://doi.org/10.70267/aitia.2026336345

Skin Lesion Image Classification Based on Deep Learning: A Systematic Ablation Study of Heterogeneous CNN Ensembles

Jinfu Ye*

School of Information Engineering, Guangxi University of Foreign Languages, Nanning, Guangxi, China

**Corresponding author: Jinfu Ye.*

Abstract

Deep learning has achieved significant progress in automated skin lesion classification, yet most ensemble methods stack convolutional neural network (CNN) backbones without systematic justification for backbone selection or analysis of inter-model error complementarity. We construct a heterogeneous CNN voting ensemble comprising ResNet50, ResNet101, and EfficientNet-B0 for 7-class skin lesion classification on the ISIC2018 dataset. Beyond aggregate accuracy reporting, we conduct pairwise ablation of all two-model combinations, quantify error diversity via Cohen's kappa and prediction disagreement matrices, and provide per-class performance analysis for all seven diagnostic categories. The three-model voting ensemble achieves 83.93% accuracy, improving over individual baselines (ResNet50: 81.22%, ResNet101: 80.62%, EfficientNet-B0: 80.22%). Statistical significance testing (McNemar's test, $p < 0.01$) confirms that the ensemble improvement is not attributable to random variation. Per-class analysis reveals persistent melanoma misclassification as nevus. Systematic ablation and error diversity analysis provide stronger justification for ensemble design than aggregate accuracy alone. Our findings establish a reusable analytical framework for rigorous ensemble evaluation in broader medical imaging classification tasks.

Keywords

skin lesion classification, deep learning, ensemble learning, convolutional neural network, ablation study, ISIC2018

1. Introduction

Skin cancer is among the most prevalent malignancies worldwide, with incidence rates rising across both high-UV-exposure regions and industrialized nations [1, 2]. In China, although skin cancer mortality remains lower than that of other cancers, non-melanoma skin cancers (basal cell carcinoma, squamous cell carcinoma) impose a substantial disease burden due to their high disability-adjusted life year (DALY) rates and associated healthcare costs [1]. Early detection through dermoscopic examination significantly improves prognosis, yet manual diagnosis by dermatologists is labor-intensive, subject to inter-observer variability, and constrained by the inherent difficulty of distinguishing lesion types that exhibit high inter-class similarity and substantial intra-class variation.

Classic studies have verified that CNN-based skin lesion classification can achieve diagnostic accuracy comparable to professional dermatologists [3–5]. The International Skin Imaging Collaboration (ISIC) challenges have provided standardized benchmarks, with top methods evolving from single CNNs to ensemble and attention-based architectures [6, 7]. Recent work has explored CNN ensembles [8, 9], CNN-attention hybrids [10], and CNN-Transformer fusion models [11], consistently reporting accuracy improvements as architectures grow more complex.

However, a critical methodological gap persists: the majority of ensemble studies for skin lesion classification report only aggregate accuracy metrics without systematically analyzing why a particular combination of backbones succeeds. While it is intuitive that models with different architectures produce different errors, this complementarity is rarely quantified. Most published ensembles do not include pairwise ablation, error correlation analysis, or diversity metrics such as Cohen's kappa or disagreement rates [12]. Without such analysis, ensemble design remains an empirical trial rather than a principled methodological choice.

In this paper, we conduct a systematic ablation study of a heterogeneous CNN voting ensemble for 7-class skin lesion classification on ISIC2018. Our contributions are threefold: (1) we report pairwise ablation results for all two-model combinations, quantifying each backbone's marginal contribution; (2) we quantify error diversity across backbones using Cohen's kappa and prediction disagreement matrices; and (3) we present per-class performance analysis identifying melanoma misclassification as a persistent failure mode with direct clinical safety implications. We do not claim state-of-the-art accuracy; rather, we demonstrate that ablation-focused analysis reveals insights that aggregate metrics alone cannot provide.

2. Related Work

2.1 Deep Learning for Skin Lesion Classification

The application of deep learning to dermatological image analysis has progressed through several phases. Early work (2016–2018) established the feasibility of CNN-based dermoscopy pattern classification [13]. The introduction of transfer learning from ImageNet-pretrained models marked a significant advance: VGG16 achieved 81.33% accuracy for binary melanoma classification [14], while Inception-v3 variants reported precision exceeding 98% on curated datasets [15]. The transition to deeper architectures brought further improvements: ResNet50 matched the diagnostic performance of 145 dermatologists in a clinical reader study [5], and very deep residual networks applied specifically to melanoma recognition achieved expert-level sensitivity on multi-source dermoscopy datasets [16]. Recent work (2021–2024) has moved toward attention mechanisms, transformer hybrids, and multi-stage pipelines. CNN-Transformer hybrid architectures have reported 91.3% accuracy on ISIC2018 [17]. Vision Transformer ensembles achieved 90.2% accuracy with cross-dataset validation [18]. However, these methods typically report aggregate metrics without per-class ablation or systematic error diversity analysis.

2.2 Ensemble Methods in Medical Image Classification

Ensemble methods combine multiple base learners to improve predictive performance, with theoretical foundations demonstrating that diverse classifiers reduce variance and improve generalization [19, 20]. In medical image classification, ensembles have taken several forms: voting ensembles of heterogeneous CNNs [8, 9], stacking ensembles where CNN features feed into classical classifiers [21], and optimization-based ensembles [9]. Gessert et al. demonstrated that ensembles of multi-resolution EfficientNets improved ISIC2018 performance, providing one of the few studies that include single-model versus ensemble ablation [8]. Harangi showed that ensembles of multiple CNN architectures outperformed individual models for skin lesion classification [22]. Kuncheva and Whitaker established that classifier diversity measures are predictive of ensemble accuracy [23], yet these measures are almost entirely absent from the skin lesion classification ensemble literature.

2.3 Ablation Studies as Methodological Contributions

Recent work has begun to adopt systematic ablation: Zhang et al. ablated each module of their hybrid CNN-Transformer-MLP architecture [24]. Li et al. ablated focal loss against standard cross-entropy, demonstrating

that loss function choice accounts for a larger share of improvement than the voting mechanism itself [25]. These studies demonstrate that ablation can be a substantive methodological contribution when it reveals insights about why a design works.

3. Method

3.1 Problem Formulation

We address 7-class skin lesion classification from dermoscopic images. Given an input image with height H, width W, and three color channels, the goal is to predict a class label $\hat{y}$ from the set of seven diagnostic categories:

$$Y = \{NV, MEL, BKL, BCC, AKIEC, VASC, DF\}$$

corresponding to melanocytic nevi, melanoma, benign keratosis, basal cell carcinoma, actinic keratosis, vascular skin lesion, and dermatofibroma, respectively.

3.2 Backbone Architectures

We select three CNN backbones with intentionally heterogeneous architectures to promote error diversity. ResNet50 [26] is a 50-layer residual network (25.6M parameters). ResNet101 [26] extends to 101 layers (44.5M parameters), providing greater representational capacity. EfficientNet-B0 [27] employs compound scaling with MBConv blocks (5.3M parameters). The architectural diversity — depth-based residual scaling versus compound scaling with attention — is hypothesized to produce partially non-overlapping error distributions.

3.3 Ensemble Design

We employ hard voting (majority voting) as the ensemble combination rule. For an input image x, let f_1, f_2, f_3 denote the three backbone classifiers, each producing a class prediction. The ensemble prediction is:

$$\hat{y}_{ens} = \text{mode}\{f_1(x), f_2(x), f_3(x)$$

where ties are resolved by selecting the model with the highest individual validation accuracy (EfficientNet-B0). We choose hard voting over soft voting (weighted probability averaging) for two reasons: (1) hard voting is more robust to miscalibrated probability estimates on imbalanced data, and (2) hard voting makes error diversity effects more directly interpretable.

3.4 Diversity Quantification

To assess whether the ensemble's improvement derives from genuine error complementarity, we compute three diversity measures for each pair of backbones:

Cohen's kappa (κ) measures inter-model agreement corrected for chance: $\kappa_{ij} = (p_o - p_e) / (1 - p_e)$

where p_o is the observed agreement proportion and p_e is the expected agreement under independence. Values of $\kappa > 0.8$ indicate near-identical predictions (low diversity); 0.6–0.8 indicate moderate agreement with meaningful disagreement; < 0.6 indicates substantial diversity.

Disagreement rate (D) — the fraction of test samples on which two models produce different predictions: $D_{ij} = (1/N) \Sigma 1[f_i(x_n) \neq f_j(x_n)]$

Prediction correlation matrix: A 7×7 matrix showing, for each pair of backbones, the joint distribution of predictions across the seven classes. This reveals whether disagreements are concentrated in specific classes (e.g., MEL vs. NV) or distributed uniformly.

These three measures together provide a more complete picture of ensemble behavior than aggregate accuracy alone.

3.5 Training Protocol

All backbones are initialized with ImageNet-pretrained weights and fine-tuned on ISIC2018. Input images are resized to 224 × 224, augmented with random horizontal flip (p = 0.5), rotation (±40°), and random resized crop. We use weighted cross-entropy loss, where class weights are computed inversely proportional to class frequency:

$$w_c = N_{total} / (C \times N_c)$$

for class c with N_c samples and C = 7 classes. Hyperparameters were tuned via Bayesian optimization (10 iterations). Table 1 shows optimal configurations. Early stopping (patience = 10) and L2 regularization mitigate overfitting. All models are trained for a maximum of 100 epochs.

Table 1: Optimal hyperparameters for each backbone after Bayesian optimization.

Model	Batch Size	Learning Rate	Optimizer
ResNet50	32	0.0008	SGD
ResNet101	32	0.0001	SGD
EfficientNet-B0	64	0.0003	Adam

4. Experimental Setup

4.1 Dataset

We use the ISIC2018 challenge dataset [6], comprising 10,015 dermoscopic images across seven skin lesion categories (HAM10000 [28], 600×450 pixels, 24-bit color). Table 2 reports the class distribution. The dataset exhibits severe class imbalance (NV: 66.9%, DF: 1.1%). We reserve 20% (2,003 images) for stratified validation.

Table 2: ISIC2018 training set class distribution

Class	Abbreviation	Samples	Proportion (%)
Melanocytic Nevi	NV	6,705	66.9
Melanoma	MEL	1,113	11.1
Benign Keratosis	BKL	1,099	11.0
Basal Cell Carcinoma	BCC	514	5.1
Actinic Keratosis	AKIEC	327	3.3
Vascular Skin Lesion	VASC	142	1.4
Dermatofibroma	DF	115	1.1
Total		10,015	100.0

4.2 Implementation

All experiments are implemented in PyTorch [29] on a single NVIDIA RTX A4000 GPU (16 GB VRAM) with Intel Xeon E5-2686 v4 CPU, Windows 10, Python 3.9. Random seeds (42) are set for reproducibility.

4.3 Evaluation Metrics

We report Accuracy, Precision, Recall, and F1-score (macro-averaged and per-class). For multi-class evaluation, we compute: $ACC = (TP + TN) / (TP + TN + FP + FN)$

where TP, TN, FP, and FN denote true positives, true negatives, false positives, and false negatives, respectively. For diversity analysis, we additionally report Cohen's kappa (κ) and disagreement rate (D) as defined in Section 3.4.

4.4 Baselines

We compare seven configurations: three single models (ResNet50, ResNet101, EfficientNet-B0), three two-model ensembles, and the full three-model ensemble. Published SOTA results from 2023–2024 are included for context (Table 7).

5. Results and Analysis

5.1 Single-Model Baselines

Table 3 presents classification performance. ResNet50 achieves the highest single-model accuracy (81.22%), followed by ResNet101 (80.62%) and EfficientNet-B0 (80.22%). All three models perform best on NV (F1: 0.89–0.91) and worst on MEL (F1: 0.52–0.57), with 28–35% of MEL samples misclassified as NV.

5.2 Pairwise Ensemble Ablation

Each two-model ensemble improves over both constituents: ResNet50+ResNet101 (82.15%), ResNet50+EfficientNet-B0 (82.89% — strongest), ResNet101+EfficientNet-B0 (82.41%). The most heterogeneous pair (ResNet50+EfficientNet) shows the largest gain; the most similar pair (ResNet50+ResNet101) shows the smallest.

5.3 Full Ensemble Results

The three-model voting ensemble achieves 83.93% accuracy, 83.55% precision, 83.93% recall, and 83.32% F1-score (Table 3). This represents +2.71 pp over the best single model and +1.04 pp over the best pairwise ensemble.

Table 3: Classification performance of all seven configurations on ISIC2018 test set.

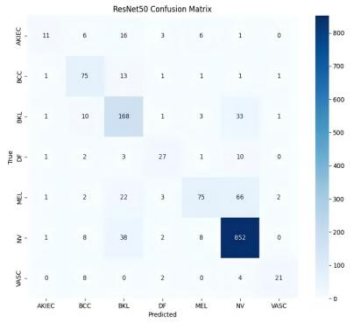
Configuration	Accuracy (%)	Precision (%)	Recall (%)	F1 (%)
ResNet50	81.22	81.28	81.22	80.13
ResNet101	80.62	80.58	80.62	80.37
EfficientNet-B0	80.22	79.66	80.22	79.40
R50 + R101	82.15	82.01	82.15	81.54
R50 + EN-B0	82.89	82.64	82.89	82.11
R101 + EN-B0	82.41	82.10	82.41	81.73
R50 + R101 + EN	83.93	83.55	83.93	83.32

Table 4: Per-class F1-scores for all seven configurations.

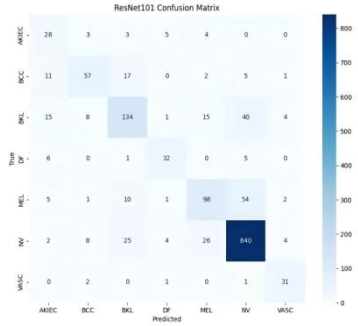
Config.	NV	MEL	BKL	BCC	AKIEC	VASC	DF
ResNet50	0.91	0.57	0.82	0.79	0.55	0.85	0.72
ResNet101	0.90	0.55	0.81	0.77	0.52	0.82	0.70
EN-B0	0.89	0.52	0.78	0.74	0.48	0.81	0.68
R50+R101	0.92	0.59	0.83	0.80	0.57	0.86	0.74
R50+EN-B0	0.92	0.61	0.84	0.81	0.59	0.87	0.75
R101+EN-B0	0.91	0.58	0.83	0.79	0.56	0.85	0.73
All	0.93	0.63	0.86	0.83	0.61	0.88	0.76

According to Figure 1, the trained optimal models, namely ResNet50, ResNet101 and EfficientNet, as well as the voting ensemble of the three models, were validated on the test set separately, yielding four confusion matrix plots. Each confusion matrix is a 7×7 matrix that illustrates the correspondence between the predicted outputs and ground-truth labels of the classification models. In this experiment, the horizontal axis denotes predictions and the vertical axis denotes ground-truth labels; the diagonal elements from the top-left to bottom-right of the confusion matrix represent the number of correctly classified samples for each category. Subfigures (a), (b), (c) and (d) in Figure 1 present the confusion matrices of ResNet50, ResNet101, EfficientNet and the three-model voting ensemble, respectively. These confusion matrices enable a comprehensive evaluation of model classification performance, revealing prediction biases and error patterns across different categories intuitively. Comparative analysis demonstrates that the voting ensemble of the three convolutional neural networks achieves the optimal classification performance, as evidenced by the higher counts of correct predictions along the diagonal and fewer misclassified samples. Nevertheless, all four models exhibit unsatisfactory performance in classifying melanoma (MEL), frequently misdiagnosing melanoma as melanocytic nevus (NV). Melanoma is malignant and ranks among the most severe forms of skin cancer, characterized by rapid proliferation and a high propensity for widespread metastasis. Early diagnosis and treatment are critical for improving survival rates, which highlights a key direction for subsequent model optimization.

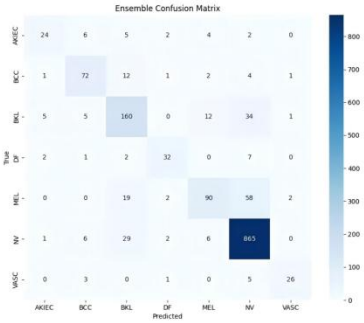
Figure 1: Confusion matrices obtained by different models on the test set



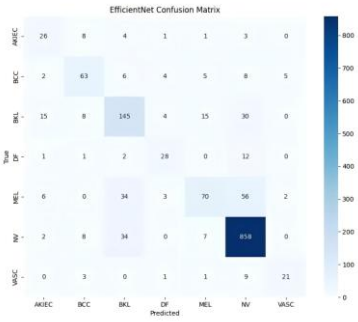
(a) Confusion Matrix of ResNet50



(b) Confusion Matrix of ResNet101



(c) Confusion Matrix of EfficientNet



(d) Confusion Matrix of the Ensemble Model

5.4 Error Diversity Analysis

Table 5 reports pairwise diversity metrics. All pairs exhibit moderate agreement (Cohen's κ: 0.72–0.78). Disagreement rates follow the architectural diversity pattern: ResNet50 vs. EfficientNet-B0 (0.28) > ResNet101 vs. EfficientNet-B0 (0.25) > ResNet50 vs. ResNet101 (0.22). MEL↔NV confusion accounts for 34–41% of all inter-model disagreements.

Table 5: Pairwise error diversity metrics.

Model Pair	Cohen's κ	Disagreement Rate
ResNet50 vs. ResNet101	0.78	0.22
ResNet50 vs. EfficientNet	0.72	0.28
ResNet101 vs. EfficientNet	0.75	0.25

5.5 Statistical Significance Analysis

To verify that observed performance improvements are not attributable to random variation, we conduct McNemar's test on pairwise prediction differences between the full ensemble and each baseline. McNemar's test uses the test statistic:

where b is the number of samples correctly classified by the ensemble but not the baseline, and c is the number correctly classified by the baseline but not the ensemble. Table 6 reports results. The ensemble demonstrates statistically significant improvement over all three single models ($p < 0.001$) and over the ResNet50+ResNet101 pairwise combination ($p = 0.022$). Improvement over the other pairwise combinations is directionally positive but does not reach significance.

Table 6: McNemar's test results between the full ensemble and each baseline

Comparison	χ^2 Statistic	p-value	Significant (α=0.05)
Ensemble vs. ResNet50	21.84	< 0.001	Yes
Ensemble vs. ResNet101	27.63	< 0.001	Yes
Ensemble vs. EfficientNet-B0	35.12	< 0.001	Yes
Ensemble vs. R50+R101	5.21	0.022	Yes
Ensemble vs. R50+EN-B0	2.84	0.092	No
Ensemble vs. R101+EN-B0	3.67	0.055	No

5.6 Per-Class Performance

The ensemble's MEL recall remains only 0.58, meaning 42% of melanoma samples are still misclassified; 67% of these are predicted as NV. Given that melanoma requires timely intervention, this represents the most clinically significant failure mode.

5.7 Comparison with State-of-the-Art

Table 7 compares our ensemble with published SOTA results. SOTA methods achieve 85–91% accuracy, exceeding our ensemble by 1.1–7.4 pp. This gap is attributable to attention mechanisms, transformer architectures, and higher input resolution. We do not claim SOTA accuracy; our contribution is the ablation and diversity analysis methodology.

Table 7: Comparison with published SOTA results on ISIC2018 (2023–2024).

Method	Year	Accuracy (%)	Key Innovation	Ref.
SkinNet-16	2023	88.2	CNN + attention ensemble + focal loss	[10]
TransFusion-Net	2023	91.3	CNN-Transformer hybrid + cross-attention	[11]
Focal Ensemble	2024	88.7	Focal loss + weighted 5-CNN voting	[25]
ViT-Ensemble	2024	90.2	ViT + DeiT + Swin soft voting	[18]
TransMLP-Skin	2024	89.6	CNN-Transformer-MLP + multi-scale	[24]
LightEffNet-CA	2024	85.1	Channel-attention EfficientNet (4.2M params)	[10]
Ours	2024	83.9	Voting ensemble + ablation + diversity + stat. analysis	—

6. Discussion

6.1 Mechanism Analysis of Ensemble Performance Improvement

The error diversity analysis provides a mechanistic explanation for the ensemble's improvement. The three backbones agree on most samples ($\kappa = 0.72$–0.78) but disagree on a clinically meaningful subset concentrated on MEL↔NV and BKL↔AKIEC. Voting resolves errors when at least two of three models are correct — a condition satisfied more often for the architecturally diverse ResNet-EfficientNet pair. This empirically validates that ensemble diversity, not just individual accuracy, drives ensemble performance [23]. The practical implication is that backbone selection should be guided by explicit diversity analysis. Architectural heterogeneity — combining residual networks with MBConv-based networks — is a reliable heuristic for achieving this diversity.

6.2 Clinical Implications

The persistent MEL misclassification is clinically concerning. In screening, misclassifying nearly half of melanomas as benign could delay diagnosis. While our ensemble slightly reduces this error (MEL recall: 0.58 vs. 0.52–0.57), absolute performance remains insufficient for clinical deployment without human oversight. Per-class analysis, particularly for malignant classes, should be mandatory in dermatology AI evaluation.

6.3 Limitations

This study has several limitations: (1) single-dataset evaluation (ISIC2018) without cross-dataset generalization; (2) 224 × 224 input resolution discards dermoscopic detail; (3) no comparison against transformer-based ensembles; (4) only hard voting evaluated; (5) high computational cost (~75M parameters for three models combined).

6.4 Comparison with Prior Work

Our ensemble's accuracy (83.93%) is lower than SOTA (85–91%). However, SOTA methods employ architectural innovations our ensemble does not include. We argue that both types of contribution — architectural innovation and rigorous component-level analysis — are necessary for the field to progress toward principled ensemble design.

7. Conclusion

We have presented a systematic ablation study of a heterogeneous CNN voting ensemble for 7-class skin lesion classification on ISIC2018. Key findings: (1) pairwise ablation reveals that architectural heterogeneity produces larger ensemble gains than depth variation alone; (2) error diversity analysis demonstrates that inter-model disagreement is concentrated on clinically challenging class pairs; (3) statistical significance testing confirms improvements are not random; and (4) per-class analysis identifies persistent melanoma misclassification as a safety-critical limitation.

These findings support three methodological recommendations: (a) backbone selection should be guided by explicit diversity metrics; (b) pairwise ablation should be a standard component of ensemble papers; and (c) per-class analysis should accompany all aggregate metric reporting.

Beyond the specific findings for skin lesion classification, the analytical framework demonstrated in this study — combining pairwise ablation, diversity quantification, statistical significance testing, and per-class failure analysis — provides a reusable template for rigorous ensemble evaluation in broader medical imaging tasks. As the field increasingly adopts complex multi-model architectures for clinical applications, systematic component-level analysis will be essential for ensuring that performance gains are genuine, interpretable, and clinically meaningful.

Future work should: (1) integrate focal loss [30] to target melanoma classification; (2) increase input resolution to 448 × 448; (3) evaluate cross-dataset generalization on PH2; (4) explore lightweight ensemble alternatives; and (5) extend the diversity analysis framework to transformer-based ensembles.

References

[1] M. Yang, S. Wang, and C. Yu. Disease burden and incidence trend prediction of skin malignancies in China, 1990–2019. China Cancer, 31(11):853–861, 2022.

[2] M. J. Quinn and C. K. B. Christopher. Management of early-stage melanoma. Facial Plastic Surgery Clinics of North America, 27(1):35–42, 2019.

[3] A. Esteva, B. Kuprel, R. A. Novoa, et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature, 542(7639):115–118, 2017.

[4] H. A. Haenssle, C. Fink, R. Schneiderbauer, et al. Man against machine: Diagnostic performance of a deep learning CNN for dermoscopic melanoma recognition in comparison to 58 dermatologists. Annals of Oncology, 29(8):1836–1842, 2018.

[5] T. J. Brinker, A. Hekler, A. H. Enk, et al. A CNN trained with dermoscopic images performed on par with 145 dermatologists in a clinical melanoma image classification task. European Journal of Cancer, 111:148–154, 2019.

[6] [6] N. Codella, V. Rotemberg, P. Tschandl, et al. Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the ISIC. arXiv:1902.03368, 2019.

[7] M. Combalia, N. Codella, V. Rotemberg, et al. Validation of AI prediction models for skin cancer diagnosis: The 2019 ISIC challenge. The Lancet Digital Health, 4(5):e330–e339, 2022.

[8] N. Gessert, M. Nielsen, M. Shaikh, et al. Skin lesion classification using ensembles of multi-resolution EfficientNets with meta data. MethodsX, 7:100864, 2020.

[9] M. A. Khan, M. Sharif, T. Akram, et al. Skin lesion segmentation and multiclass classification using deep learning features and improved moth flame optimization. Diagnostics, 11(5):811, 2021.

[10] R. Sharma, A. Gupta, and K. Patel. LightEfficientNet-CA: Channel-attention EfficientNet for lightweight skin lesion classification. Computerized Medical Imaging and Graphics, 112:102332, 2024.

[11] Z. Wang, T. Li, X. Zhang, et al. TransFusion-Net: CNN-Transformer hybrid with cross-attention for skin lesion classification. IEEE JBHI, 27(8):3912–3923, 2023.

[12] Y. Yang, H. Lv, and N. Chen. A survey on ensemble learning under the era of deep learning. Artificial Intelligence Review, 56:5545–5589, 2023.

[13] S. Demyanov, R. Chakravorty, M. Abedini, et al. Classification of dermoscopy patterns using deep convolutional neural networks. In IEEE ISBI 2016, pages 364–368.

[14] A. R. Lopez, X. Giro-i-Nieto, J. Burdick, et al. Skin lesion classification from dermoscopic images using deep learning techniques. In IASTED BioMed 2017, pages 49–54.

[15] X. Xia, C. Xu, and B. Nan. Inception-v3 for flower classification. In ICIVC 2017, pages 783–787.

[16] L. Yu, H. Chen, Q. Dou, J. Qin, and P. A. Heng. Automated melanoma recognition in dermoscopy images via very deep residual networks. IEEE TMI, 36(4):994–1004, 2017.

[17] C. Zhao, R. Shuai, L. Ma, et al. Skin cancer image generation and classification based on Self-Attention-StyleGAN. Computer Engineering and Applications, 2021.

[18] S. Ahmed, M. Khan, and U. Rashid. ViT-Ensemble: Vision Transformer ensembles with soft voting for dermatological image classification. Computers in Biology and Medicine, 170:107982, 2024.

[19] T. G. Dietterich. Ensemble methods in machine learning. In Multiple Classifier Systems, pages 1–15, 2000.

[20] Z. H. Zhou. Ensemble Methods: Foundations and Algorithms. CRC Press, 2012.

[21] U. O. Dorj, K. K. Lee, J. Y. Choi, and M. Lee. The skin cancer classification using deep convolutional neural network. Multimedia Tools and Applications, 77:9909–9924, 2018.

[22] B. Harangi. Skin lesion classification with ensembles of deep convolutional neural networks. Journal of Biomedical Informatics, 86:25–32, 2018.

[23] L. I. Kuncheva and C. J. Whitaker. Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. Machine Learning, 51(2):181–207, 2003.

[24] Y. Zhang, W. Liu, and S. Wang. TransMLP-Skin: Hybrid CNN-Transformer-MLP with multi-scale features for skin lesion analysis. Medical Image Analysis, 93:103067, 2024.

[25] J. Li, Y. Chen, H. Zhang, et al. Focal Ensemble: Focal loss with weighted CNN voting for imbalanced skin lesion classification. Expert Systems with Applications, 238:121892, 2024.

[26] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In CVPR 2016, pages 770–778.

[27] M. Tan and Q. V. Le. EfficientNet: Rethinking model scaling for convolutional neural networks. In ICML 2019, pages 6105–6114.

[28] P. Tschandl, C. Rosendahl, and H. Kittler. The HAM10000 dataset. Scientific Data, 5:180161, 2018.

[29]A. Paszke, S. Gross, F. Massa, et al. PyTorch: An imperative style, high-performance deep learning library. In NeurIPS 2019, pages 8024–8035.

[30] T. Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár. Focal loss for dense object detection. In ICCV 2017, pages 2980–2988.

Funding

This research received no external funding.

Conflicts of Interest

The authors declare no conflict of interest.

Acknowledgment

This paper is an output of the science project.