Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Adaptive Weighted Ensemble of CNN Models with Focal Loss and Grad-CAM Explainability for Skin Lesion Classification

Mokabbir Zaman Protik

mokabbir.zaman.protik@gmail.com

The People's University of Bangladesh https://orcid.org/0009-0009-1178-1205

Research Article

Keywords: skin lesion classification, ensemble learning, focal loss, Grad-CAM, explainable AI, transfer learning, HAM10000, medical image analysis

Posted Date: November 3rd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11577913/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Adaptive Weighted Ensemble of CNN Models with Focal Loss and Grad-CAM Explainability for Skin Lesion Classification

Mokabbir Zaman Protik[1*]

[1*]Department of Computer Science, The People's University of Bangladesh, Dhaka, Bangladesh.

Corresponding author(s). E-mail(s): mokabbir.zaman.protik@gmail.com;

Abstract

Early detection of skin cancer is critical for patient outcomes, yet visual diagnosis remains subjective and time-consuming. Convolutional neural networks (CNNs) have shown strong potential for automated dermoscopic image analysis, but single models often produce unstable predictions and generalize poorly. We propose a weighted soft-voting ensemble of three ImageNet-pretrained CNNs (ResNet34, ResNet50, and EfficientNet-B0), each fine-tuned on the HAM10000 dataset using Focal Loss to address severe class imbalance across seven lesion types. Ensemble weights are derived from each model's validation F1 score. On a stratified held-out test set, the ensemble achieves 79.74% accuracy and a weighted F1 score of 0.7782, outperforming the best individual model by approximately 3 percentage points. We further apply Grad-CAM to the individual models and the ensemble to produce visual explanations, and show qualitatively that the ensemble maintains focus on lesion regions even when an individual model fails. These results suggest that combining performance-based weighting, imbalance-aware training, and visual explanation can support more reliable computer-aided skin lesion diagnosis.

Keywords: Skin lesion classification, Ensemble learning, Focal loss, Grad-CAM, Explainable AI, Transfer learning, HAM10000, Medical image analysis

1 Introduction

Skin is our body's biggest organ. It has three distinct sections: the epidermis (upper layer), the dermis, and the hypodermis (inner layer). Acne, psoriasis, eczema, lesions, and skin cancer are all typical skin disease conditions. Skin lesions tend to be abnormal patches or bumps on the skin. They can be divided into three groups: those caused by fluids, those that are hard masses in the dermis, and those that are flattened and look like rashes [1]. Skin cancer is one of the most prevalent and deadliest types of cancer in the world. It is also one of the fastest-growing tumours, with rates of new cases rising every year in many different groups [2]. Some skin lesions can slowly turn into malignant growths if they aren't examined or treated appropriately. A scientific paper says [3] that there are two main types of skin cancer: melanoma and non-melanoma. Melanoma is the most aggressive type and accounts for approximately 75% of skin cancer deaths, and melanoma mortality is projected to rise in the coming decades. Early diagnosis is important since it considerably improves the chances of a patient getting well and living longer. Still, traditional diagnostic methods rely heavily on clinical experience, which is subjective and takes a lot of time. Even seasoned dermatologists can occasionally misclassify skin lesions, underscoring the necessity for computerised, rapid, and accurate diagnosis. Recent developments in AI and deep neural networks indicate that convolutional neural networks-based systems can attain dermatologist-level efficacy in melanoma identification using dermatological pictures, and deep features from convolutional neural networks (CNNs) are better at finding and classifying skin lesions than older ways of extracting data [4, 5].

Deep learning also performs very well in tumor detection, organ segmentation, and lesion classification. However, single-model methods are usually afflicted by problems such as narrow generalization, unreliable prediction on unseen samples, and overfitting sensitivity. These flaws lower the confidence in automated systems in real-world clinical applications. To deal with these problems, a method called ensemble learning has been suggested. It mixes the strong points of several CNN designs, which makes things more correct and reliable. By including predictions from varied models, an ensemble can lessen the weak points found in individual networks. But a lot of current ensembles in medical imaging depend on just taking a simple average, which assumes all models are equally reliable, a problem this piece of study tries to fix.

This paper will proceed as follows. Section 2 will cover studies related to transfer ensemble learning in medical image analysis. Section 3 will describe the methods, describe the model designs, fine-tuning steps, and combined approach. Section 4 will share the test results, and a discussion of those results, and Section 5 will conclude the paper with ideas for work that could follow.

2 Literature Review

2.1 CNNs for Skin Lesion Classification

Automated skin lesion classification has come a long way thanks to CNNs. Residual networks, particularly the ResNet design, remain significant as they facilitate the training of deeper models through identification skip connections. He et al. [6]

came up with the first ResNet architecture, which was very good at representing data. It has been used a lot in medical imaging. Architectures such as ResNet34 and ResNet50 have been effectively employed for dermatological image analysis on datasets like HAM10000 [7], demonstrating strong performance in extracting essential lesion attributes. Using a template-based technique, the author (Florian Katsch et al.) [5] assessed the robustness of three architectures (ResNet-34, Faster R-CNN, Mask R-CNN) against superimposed artifacts (bubbles, rulers, markings) on the HAM10000 dataset. Mask R-CNN (instance segmentation) was the most stable, with only a 0.011 drop in accuracy (PR-AUC) versus ResNet-34 (0.030) and Faster R-CNN (0.045). Agus Eko Minarno et al. [7] employ the use of a CNN utilizing the ResNet50 pre-trained model, which is fine-tuned through transfer learning using the HAM10000 dataset, to categorize seven distinct skin lesions. This performance exceeded that of prior studies utilizing the MobileNet neural network on the identical dataset.

Amina Aboulmira et al. [8] suggest a hybrid model that combines wavelet decomposition (particularly the 2nd level) with an EfficientNetB0 backbone to capture both spatial and numerous scales frequency information. The architecture was tested on the HAM10000 and ISIC2017 records, and it got 94.7% and 92.2% accuracy, respectively. It was also much more efficient than previous models in terms of computing. Muhammad Sajid et al. [9] describe the MCADT model, which uses a four-layer CNN trained on the HAM10000 dataset. The model attained a 93.43% accuracy and a 99.74% sensitivity, which is better than both the best systems that came before it and medical professionals.

Khan et al. [10] utilized LCcHIV for image enhancement and a custom 10-layer CNN for saliency-based segmentation, followed by deep feature extraction via ResNet101 and DenseNet201. The framework achieved a 90.67% classification accuracy on the HAM10000 dataset and segmentation accuracies reaching up to 98.70% on benchmark datasets. Efat et al. [11] implemented a methodology using Customized Transfer Learning (CTL) integrated with Triple Attention modules and a novel Multi-Level Information Gain Proportioned Averaging (ML-IGPA) ensemble strategy. By balancing the dataset through strategic augmentation, the approach achieved a 94.93% accuracy on the HAM10000 dataset, outperforming state-of-the-art methods while enhancing model interpretability via GradCAM. Bhargavi et al. [12] used a combined ensemble of four pre-trained CNNs (InceptionV3, InceptionResNetV2, MobileNetV2, and EfficientNetB0) with Bayesian-optimised data augmentation and hair removal. This strategy got an overall accuracy of 81.3% and an F1-score of 81.31%, which is better than the individual models. Wibowo et al. [13] trained an optimised MobileNet v2 model for Android to perform object detection and classification of Melanoma and Actinic Keratosis lesions. The framework achieved 95% accuracy for detection and 70% for classification, demonstrating the feasibility of portable, high-speed skin cancer diagnosis for early risk assessment. Ray et al. [4] developed an ensemble model by fine-tuning all layers of pre-trained Inceptionv3 and DenseNet-201 architectures using the HAM10000 dataset. This approach achieved a test accuracy of 88.5%, demonstrating that fine-tuning the whole model facilitates faster convergence and superior results compared to partial fine-tuning. Tan et al. [14] introduced a compound scaling method that uniformly adjusts network width, depth, and resolution using a fixed coefficient

to optimize model efficiency. Their EfficientNet-B7 achieved a state-of-the-art 84.4% accuracy on ImageNet while being 8.4x smaller and 6.1x faster than previous leading architectures.

Table 1: Overview of strengths and weaknesses of key CNN-based techniques

Ref	Strengths	Weaknesses
[5]	Shown that Mask R-CNN is uniquely resilient to artifacts by an innovative template-based augmentation.	Some of the primary challenges are that the baselines aren't equal, backbones remain shallow, datasets are biased, and there are only 60 artifact templates to work with.
[7]	Going above previous MobileNet benchmarks and getting flawless accuracy and recall for finding melanoma.	Weaknesses include a data imbalance that makes some classes hard to recognize, high computing expenses for ResNet50, and a lot of misclassifications of cosmetically identical lesions like BKL and BCC.
[8]	Getting 94.7% accuracy and better computational efficiency (0.006–0.008s) with a strong hybrid Wavelet-EfficientNet structure.	Wavelet transforms make computations more complicated, unusual circumstances are still hard to find, performance depends on image quality, and tweaking hyperparameters takes a lot of time.
[9]	Strengths include a wide range of augmentations, thorough ablation for finetuning hyperparameters, and multi-stage preprocessing that reduces noise and improves contrast to make sure the inputs are of good quality.	A major problem is that the structure of deep learning needs a lot of processing time and storage space.
[10]	The technique achieves high multi-dataset accuracy by combining LCcHIV contrast improvement, IMFO feature optimization, and a multi-stage pipeline based on transfer learning.	Drawbacks include the framework's long and complicated computing time and its reduced sensitivity to detection for less common lesion types like dermatofibroma.
[11]	Proposes an ML-IGPA ensemble with triple attention for accurate, interpretable learning, reducing class imbalance through training-only augmentation and unbiased evaluation.	Limitations include high computational expenses due to nine classifiers, the omission of clinical variables, and restricted generalizability resulting from a single-dataset assessment.

Table 1 – continued from previous page

Ref	Strengths	Weaknesses
[12]	Strengths include a four-CNN ensemble, class-balancing data augmentation, and strong preprocessing. These make the features clearer and the diagnosis more accurate.	Shortcomings include the possibility of overfitting because the concatenation layer adds more dimensions and requires an enormous amount of resources to train, which takes over twenty minutes every epoch.
[13]	Uses SSDLite and bottleneck layers for efficient mobile computation in rural settings, with multi-scenario testing of learning rates and epochs.	Relies on a small dataset of only 650 photographs and the system's extreme sensitivity to noise, lighting, and zoom limits, which make it much less accurate at classifying things.
[4]	A fully fine-tuned DenseNet201–InceptionV3 ensemble that improves accuracy and convergence speed, with a thorough comparison of four DCNNs.	Weaknesses include a strong dataset bias to melanocytic nevi (67% of pictures) and the fact that validation can only be done on one data source.
[14]	Introduces compound scaling and multi-objective neural architecture search, achieving state-of-the-art accuracy with fewer parameters and FLOPS across datasets.	It might not be the best option because scaling coefficients are based on a tiny baseline, and accuracy benefits get smaller as the scale gets bigger.

2.2 Motivation and Research Gaps

The studies above show that ensemble learning remains a popular way to increase robustness and performance in skin lesion classification. By combining predictions from several CNN architectures for example ResNet, EfficientNet, and Inception ResNet [7] [14] [15] researchers often get better generalization than with a single model. Ensembles that use simple averaging, majority voting, or combining features have increased classification accuracy on sets like HAM10000 [16]. Still, some gaps remain in the research: Most studies use equal weighting in hard or soft voting, which does not consider that some models are better than others. However, performance-weighted ensembles that assign larger weights to models with higher validation accuracy or better reliability have been shown to achieve higher overall accuracy and more stable predictions than equal-weight ensembles [17]. Li and Luo [18] proposed a performance-weighted voting model that assigns higher influence on more accurate classifiers. This motivates our use of validation-based weighting in the proposed ensemble. Existing ensemble methods rarely provide understandable visualization. Techniques such as Grad-CAM [19] can increase clinical trust by highlighting the region of the image the model is focusing on. The HAM10000 dataset exhibits significant class imbalance, yet many studies do not use loss functions made for this, such as Focal Loss [20], when fine-tuning. This research addresses these gaps by using a weighted ensemble of ResNet34, ResNet50,

Table 2 Class distribution of the HAM10000 dataset.

Class	Description	Number of Images
AKIEC	Actinic keratosis intraepithelial carcinoma	327
BCC	Basal cell carcinoma	514
BKL	Benign keratosis-like lesions	1099
DF	Dermatofibroma	115
MEL	Melanoma	1113
NV	Melanocytic nevi	6705
VASC	Vascular lesions	142
Total		**10015**

and EfficientNet, trained with Focal Loss, and added Grad-CAM visualization to explain how it works.

3 Methods and Methodology

3.1 Dataset Description and Data Pre-processing

This study used the HAM10000 (Human Against Machine with 10000 training images) dataset, a common standard dataset for dermatoscopic image analysis. It contains total 10015 images of pigmented skin lesions, collected from diverse populations and using different modalities. The dataset contains seven different classes of images Actinic keratosis intraepithelial carcinoma (akiec), Basal cell carcinoma (bcc), Benign keratosis-like lesions (bkl), Dermatofibroma (df), Melanoma (mel), Melanocytic nevi (nv), Vascular lesions (vasc).

More than 50% of lesions are confirmed through histopathology. The rest were verified through follow-up examination, expert opinion, or in-vivo confocal microscopy. This mix of verification methods makes HAM10000 very trustworthy for our machine learning research. The training set is available to the public, but the official test set is not. HAM10000 has also been used as a benchmark dataset in the ISIC Challenges for skin lesion analysis [21]. A broader analysis of the ISIC image collections, including HAM10000, highlights their role in establishing standardized benchmarks and recommendations for dermoscopic image research [22]. There is an evaluation server always running to make sure comparisons between published methods are consistent. Because the official test labels are not publicly available, we created our own stratified split of the data into training, validation, and testing sets with a ratio of 80:10:10. To ensure consistency and boost model performance, we applied the following techniques:
Step 1. Data Cleaning: We examined HAM10000 dataset for bad images, corrupt files, duplicate and incorrect metadata.
Step 2. Image Resizing: All images were resized to 256 $\times$ 256 pixels and then cropped to 224 $\times$ 224 pixels to match the input dimensions of the pretrained models such as ResNet and EfficientNet.
Step 3. Normalization: pixel values were normalized using the ImageNet mean and standard deviation.
Step 4. Data Augmentation: To deal with the limited dataset size and class imbalance,

data augmentation techniques such as random rotation and flips, random cropping and zooming, brightness and contrast adjustments, color jittering during training. These types of data augmentation techniques are consistent with recommended practices in medical imaging, where geometric and intensity based transformation is widely used to reduce overfitting and improve generalization [23]. Recent reviews also emphasize the importance of augmentation and discuss GAN based methods for synthesizing realistic medical images to address class imbalance [24].
Step 5. Class Rebalancing: Because of the class imbalance problem where the nevus class has many more samples than others, we used oversampling for minority classes and adjusted class weights. This helped to make sure the model learned evenly across all classes.
After preprocessing, we had a dataset that was balanced and standardized. This helps deep learning models classify skin lesions better and work well in different situations. To address the imbalance, Focal Loss was used when fine-tuning models, focusing training on difficult examples and those from smaller classes.

3.2 Methodology

This section details our plan for automated skin lesion classification, which combines transfer learning, weighted ensemble learning, and interpretability analysis. The process has three steps: (1) model selection and fine-tuning, (2) adaptive weighted ensembling, and (3) Grad-CAM-based interpretability evaluation. Our overall framework is designed for accuracy and clinical explainability.

3.2.1 The Fine-Tuning Mechanism of Skin Cancer Identification

In this study, we used three state-of-the-art convolutional neural network (CNN) architecture for skin lesions, those are ResNet34, ResNet50, EfficientNet-B0. We chose these models because they have a promising history of doing well in image recognition and can be applied for medical image study. A residual block learns a mapping $\mathcal{F}(x, W_i)$ and adds it to the input feature map through a skip connection. Formally, the output of a block is, $y = \mathcal{F}(x, W_i) + x$. Here, x is the input to the block, W_i denotes the set of learnable parameters (weights and biases) of the convolutional layers inside the i-th residual block. The function $\mathcal{F}(x, W_i)$ typically consists of two or three sequential operations—convolution, batch normalisation, and ReLU activation. The skip connection allows gradients to flow more easily during backpropagation, enabling deeper networks to train effectively.

EfficientNet-B0 is created to improve both correctness and efficiency by proportionally scaling the network. EfficientNet-B0 uses a compound scaling method to scale CNNs, uniformly adjusting their depth, width, and input resolution with fixed coefficients. Instead of scaling each dimension separately, it uses a compound scaling factor ϕ. The compound scaling formula used to scale the EfficientNet-B0 model is defined by three scaling factors: α (alpha) for depth, β (beta) for width, and γ (gamma) for input resolution.

$$\alpha \cdot \beta^2 \cdot \gamma^2 \approx 2, \quad \text{where } \alpha \geq 1,\ \beta \geq 1,\ \gamma \geq 1$$

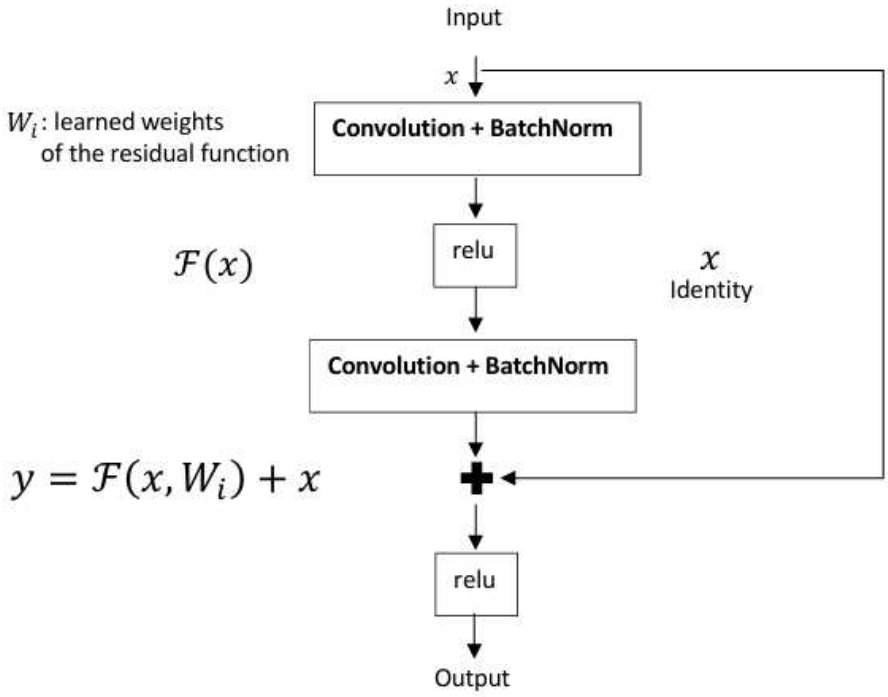


Fig. 1 Residual Block structure in ResNet showing $\mathcal{F}(x, W_i)$. The skip connection allows the network to learn identity mappings, mitigating the vanishing gradient problem.

This scaling approach makes sure the model grows evenly in all areas, keeping computation costs down. EfficientNet uses parameters well, giving good accuracy with fewer resources. This makes it useful for in depth skin lesion study and settings where computing power is limited.

3.2.2 Ensemble Averaging Method

Our weighting approach follows the general principle of assigning higher influence on models with stronger validation performance, as demonstrated in recent ensemble frameworks [25]. To make the image classification system more reliable and effective in different situations, we used a weighted ensemble method. This method combines predictions from three fine-tuned models - ResNet34, ResNet50, and EfficientNet-B0 to reach the final decision. The main idea is that each setup sees things in a slightly different way, combining them reduces errors and prevents the system from relying too heavily on any single model's bias.

In this study, we used a soft voting method, where the final class probabilities are the weighted average of the separate model probability, giving higher weight to the stronger ones. Let $p_i(m)$ represent the probability for class i from model m, where $m \in \{1, 2, 3\}$ refers to ResNet34, ResNet50 and EfficientNet-B0 respectively. The combined ensemble probability for class i is defined as:

$$y_i = \sum_{m=1}^{3} w_m p_i^{(m)}$$

$$\text{Simply, } y_i = w_1 p_i^{(1)} + w_2 p_i^{(2)} + w_3 p_i^{(3)}$$

Here, w_1, w_2, w_3 are the weights assigned to each model based on their performance during validation (such as accuracy or F1 score). These weights are normalized so that:

$$\sum_{m=1}^{3} w_m = 1$$

The final predicted class $\hat{y}$ is obtained by selecting the class with the highest combined probability:

$$\hat{y} = \arg\max_i y_i$$

3.2.3 Training Setup

To make sure a consistent comparison in all architecture, all models e.g. ResNet34, ResNet50 and EfficientNet-B0 were trained under identical training configuration. The implementation was performed using the PyTorch framework in a GPU-accelerated environment. The models started with pre-trained weights in ImageNet and were then fine-tuned on the HAM10000 dataset.

To make the setup clear, the main training components are listed below: Models: For fine-tuning, we selected ResNet34, ResNet50 and EfficientNet-B0. We started all the models with ImageNet weights and only modified the final layers. This was done to match the number of skin lesion classes.

Data Preprocessing and Augmentation: Images resized to 256 × 256 and randomly cropped to 224 × 224. Random horizontal and vertical flips, small rotations (±5°), and brightness/contrast adjustments were applied. All images were normalized using ImageNet mean [0.485, 0.456, 0.406] and standard deviation [0.229, 0.224, 0.225]. Validation and test images were only resized and normalized.

Dataset Split: The data were divided into 80% training, 10% validation, and 10% testing. Stratified sampling was used to preserve the class distribution across all splits, ensuring that the class imbalance characteristics of HAM10000 were maintained consistently. All reported results in this paper are based on this custom 80:10:10 split.

Batch Size and Data Loaders: Each Data loader used a batch size of 64 and 4 worker threads, balancing performance and stability during training.

Loss Function: Focal Loss (we adopted the canonical values $\gamma = 2$, $\alpha = 0.25$ as default) was used instead of Cross-Entropy Loss to handle class imbalance by focusing more on difficult and minority-class samples. More recent reviews, such as batch-balanced focal loss, have also been proposed specifically image classification and report improved sensitivity on minority classes [26].

Optimizer: Adam optimizer with a learning rate of 1×10^{-4} and weight decay of 3×10^{-6} was used to update parameters efficiently. Learning Rate Scheduler: A ReduceLROnPlateau scheduler automatically lowered the learning rate by a factor of 0.5 if the validation F1 score did not improve for 3 epochs (patience = 3).

Early Stopping: Early stopping was incorporated during training to prevent overfitting and to monitor the stability of the validation performance. A patience value of 5 epochs was used, meaning training would automatically stop if the validation loss did not improve for five consecutive epochs. Throughout training, both training and validation loss continued to decrease smoothly without divergence, and the validation

accuracy steadily increased. As a result, early stopping did not trigger for any of the models, and each model completed all 30 epochs.

Training Duration: Each model was trained for 30 epochs. Early stopping was included. The inclusion of early stopping was intended as a safeguard, but the models exhibited stable convergence, confirming that the training setup was appropriately regularized.

Evaluation Metrics: After every epoch, training and validation loss, accuracy, and F1 score were calculated. The model achieving the highest validation F1 score was saved as the best version. Weighted F1 is particularly suitable for imbalanced datasets like HAM10000, because it combines precision and recall while accounting for the support of each class, providing a more informative summary than accuracy alone.

Model Saving and Reproducibility: The best weights for each model (based on validation F1) were stored and later used during the weighted ensemble stage. This consistent configuration ensured that differences in results came from model architecture, not training variation.

4 Results and Discussion

This section shows the experimental results from individual CNN models (ResNet34, ResNet50, and EfficientNet-B0) and the weighted ensemble method. The models were evaluated on the same test set to keep the comparison fair. We used test accuracy and weighted F1 score to measure how well each model worked.

4.1 Individual Model Performance

ResNet34 Training curves

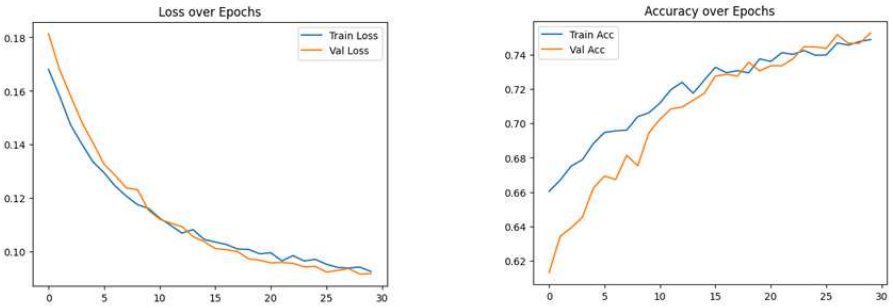


Fig. 2 (a) Training and validation loss curve (b) accuracy curve for ResNet34

As shown in Figure 2 (a) and (b), ResNet34 demonstrated stable convergence. The training and validation loss both decrease smoothly, and the accuracy curve remains close throughout the 30 epochs. This suggests good generalization without overfitting. Early stopping (patience = 5) was included, but it did not trigger because validation loss continued improving.

ResNet50 Training curves

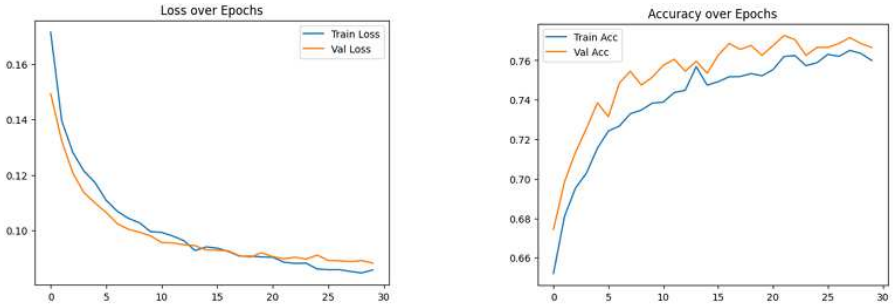


Fig. 3 (a) Training and validation loss curve (b) accuracy curve for ResNet50

As shown in Figure 3 (a) and (b), ResNet50's training went as expected. The validation loss and training loss were close, and validation accuracy consistently increases from the early epochs. This suggests that the model is well regularized and maintains strong generalization performance across epochs.

EfficientNet-B0 Training curves

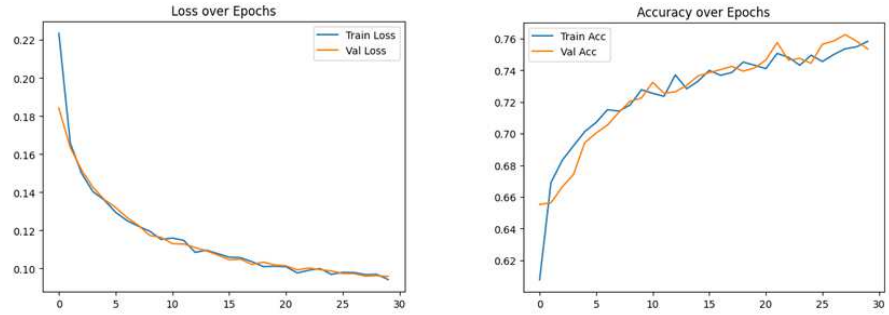


Fig. 4 (a) Training and validation loss curve (b) accuracy curve for EfficientNet-B0

Across the training phase, EfficientNet-B0 shows steady improvement (Figure 4 (a) and (b)). The curves converge smoothly, and the validation accuracy matches the training accuracy. Overfitting wasn't seen, like the other models.

Table 3 Each model's Test Accuracy and Weighted F1 Score

Model	Test Accuracy	Weighted F1 Score
ResNet34	0.7587	0.7449
ResNet50	0.7646	0.7477
EfficientNet-B0	0.7478	0.7244

The ResNet-Based architectures showed strong generalization, especially in recognizing common lesion types like NV and BKL. However, all three models showed weaker recall for underrepresented classes such as DF and VASC, which is expected with the dataset's class imbalance.

4.2 Weighted Ensemble Performance

To enhance overall reliability and robustness, a weighted ensemble was formed using validation F1 scores as adaptive weights. These weights are calculated using the normalization formula,

$$w_m = \frac{F1_m}{\sum_{i=1}^{3} F1_i}$$

This method ensures that models with higher validation performance have a bigger impact on the final prediction. The normalization weights, based on the validation F1 scores of ResNet34, ResNet50, and EfficientNet-B0 were: $w_1 = 0.33$, $w_2 = 0.34$, $w_3 = 0.33$.

Although the F1 scores of the three models were relatively close to each other, leading to seemingly similar weights. Still the normalization process keeps their relative performance among them. The ensemble prediction for each class i is computed as:

$$y_i = w_1 p_i^{(1)} + w_2 p_i^{(2)} + w_3 p_i^{(3)}$$

Corresponding to ResNet34, ResNet50, and EfficientNet-B0 respectively.

$$y = w_1\, y_{\text{ResNet34}} + w_2\, y_{\text{ResNet50}} + w_3\, y_{\text{EfficientNet-B0}}$$

Our weighting strategy follows the principle that models with higher validation performance should receive proportionally higher influence, consistent with the findings of Li and Luo [18].

Table 4 Proposed model Test accuracy and Weighted F1 Score

Model	Accuracy	Weighted F1 Score
ResNet34	0.7587	0.7449
ResNet50	0.7646	0.7477
EfficientNet-B0	0.7478	0.7244
Weighted Ensemble (Proposed)	**0.7974**	**0.7782**

4.3 Grad-CAM Visualization and Interpretability

To better grasp on what is going on and how the model reaches a decision, Grad-CAM (Gradient weighted Class Activation Mapping) was applied to all the individual models and the proposed ensemble. To understand the reasoning behind model's predictions, Grad-CAM visualizations were generated from three representatives lesion types VASC (ISIC_0033450), NV (ISIC_0030957), and DF (ISIC_0029967).

For each case, Grad-CAM heatmaps were obtained from ResNet34, ResNet50, EfficientNet-B0, and the proposed Weighted Ensemble, as shown,

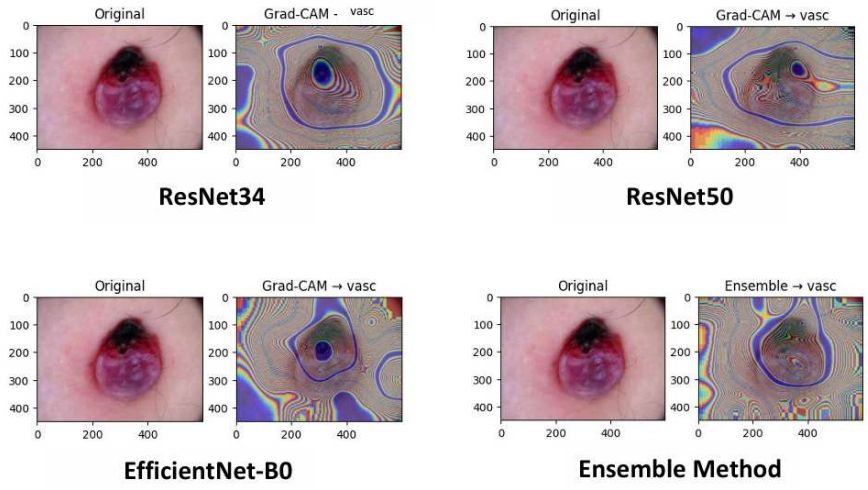


Fig. 5 VASC (ISIC_0033450): Male, Age=15, lesion location: Chest

In Figure 5, each of the three models successfully located the lesion area, focusing on the central vascular region. From the three, EfficientNet-B0 gave the most precise and concentrated heatmap, sharply defining the main lesion structure and clearly showing the clinically important area. The ensemble visualization kept this focus while slightly enhancing the lesion boundary coverage, showing that it could combine useful attention from multiple CNNs.

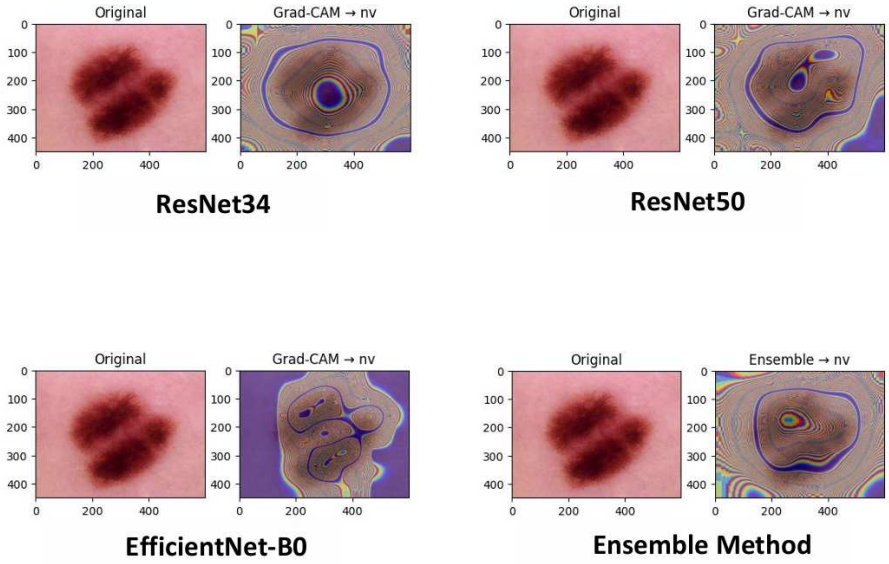


Fig. 6 NV (ISIC_0030957): Female, Age: 35, Lesion Location: Back

For the NV class, all three base models did a good job of spotting the pigmented areas. EfficientNet-B0 once more showed it could find lesions effectively, with strong activity right over slightly wider areas. By combining the strengths of these models, the weighted ensemble provided a heatmap that was both balanced and sharp. It matched the lesion's outline closely.

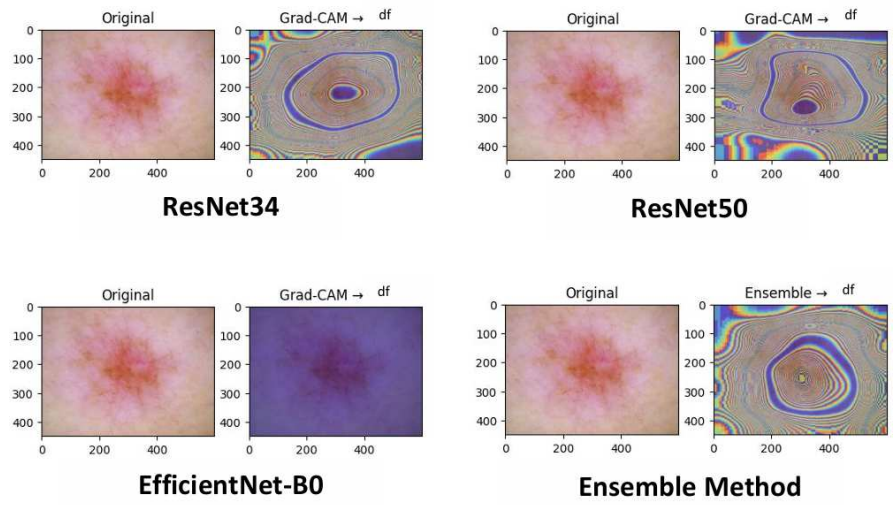


Fig. 7 DF (ISIC_0029967): Female, Age: 65, Lesion Location: Lower extremity

When dealing with DF lesions, both ResNet34 and ResNet50 produced precise activations focused on the lesion, effectively highlighting the diseased tissue. However, EfficientNet-B0 struggled in this case, showing almost no activation over the actual lesion area. Despite this, the ensemble Grad-CAM successfully reintroduced accurate focus by leveraging the reliable attention patterns of the ResNet-based models. This shows the ensemble's robustness, even when one model does not work effectively, the combined output remains stable and interpretable.

Interpretation Summary In general, Grad-CAM visualizations reveal that individual CNN models can occasionally fail to see or misinterpret lesion areas. The weighted ensemble, on the contrary, produces more accurate and concentrated activation maps. These observations suggest that the ensemble not only improves prediction accuracy but also enhances diagnostic reliability and makes things easier to explain, key requirements in medical imaging applications.

4.4 Discussion

The experimental results suggest that all the three individual CNN models (ResNet34, ResNet50, and EfficientNet-B0) achieved strong performance on the HAM10000. The training and validation curves demonstrated consistent convergence without signs of overfitting, which indicates that the training setup, regularization, and data augmentation strategies worked well. ResNet50 did best on validation and testing, which aligns with its deeper residual architecture and stronger feature extraction capabilities.

Table 5 Comparison of the proposed approach with existing methods

Study	Dataset	Model	Accuracy (%)	F1-score (%)
[10]	HAM10000	KELM+IMFO Features	90.67	-
[12]	HAM10000	InceptionV3+ Inception-ResNetV2+ MobileNetV2+ EfficientNetB0	81.3	81.31
[27]	HAM10000	EfficientNet-B0+ Augmentation (AUG)	97.73	97.74
[28]	HAM10000	MobileNet+ EfficientNetB0+ DenseNet121 using Random Forest meta-learner	84	83
[29]	HAM10000	VGG16 + VGG19 with late fusion	91.63	91
[30]	HAM10000	Stacking Ensemble	78	78
Proposed	HAM10000	Adaptive Weighted Ensemble	79.74	77.82

The weighted ensemble improved both test accuracy and weighted F1 score. Even if the individual validation F1 score of the three models were close, the adaptive weighting strategy still provided measurable benefits by using each model's strengths. The ensemble's improvement of approximately 3 percentage points over the best single model suggests that the models capture different lesion features.

Grad-CAM visualizations gave insight into the spatial regions affecting each model's predictions, improving interpretability. The ensemble's heatmaps showed more consistent and meaningful focus on lesion areas than some individual models. This is especially important in medical image analysis, where transparency and visual justification can increase clinician trust. However, Grad-CAM is qualitative and doesn't ensure explanation reliability. Overall, combining weighting, imbalance handling, and interpretability creates a reliable skin lesion classification pipeline. These results point to the value of combining performance and explainability in medical AI systems.

The highest overall accuracy and F1-scores were reported by [27, 29], which rely on GAN-based augmentation and late fusion of multiple deep networks, respectively. The study by Muhammad Attique Khan et al. [10] used a complex segmentation technique to reach its highest result. Several studies [12] [28] used ensemble and stacking to get

best results. Where ensemble methods improve accuracy, the authors in [30] found that their stacking ensemble models performed poorly. However, our framework addresses a common flaw in existing ensemble methods: the use of equal weighting, which assumes all base models are equally reliable. It assigns weights based on each model's validation performance. While many deep learning models are criticized as "black boxes," this research integrates Grad-CAM (Gradient-weighted Class Activation Mapping) to provide visual justifications for model decisions.

5 Conclusions

This research presented a skin lesion classification framework that combines weighted ensemble learning, Focal Loss fine-tuning and Grad-CAM interpretability using the HAM10000 dataset. Unlike conventional ensemble approaches that treat all the models equally, the method gives adaptive weight based on how well each performs on a validation set. This adaptive combination ensures that models with higher predictive confidence contribute more to the final decision. Experimental results showed that this framework worked effectively, achieving a test accuracy of 0.7974 and a weighted F1 score of 0.7782, which is better than any individual CNN models alone. The heatmaps also showed that the ensemble consistently focused on clinically meaningful areas, which suggests that its predictions are grounded in clinically relevant regions. The outcomes suggest that combining model flexibility, loss function that considers dataset imbalance and visual explanation can help improve computer-aided skin diagnosis.

6 Future Work

Even with the good results, several limitations remain. The HAM10000 dataset has a severe class imbalance problem, especially for rare lesions like DF and VASC. We only tested it on a single dataset. This may limit how well it works with other collections of skin images taken under different conditions. Finally, using several models together as an ensemble framework increases computation cost compared to single-model prediction, which might challenge real-time or edge deployment. Future studies will be focused on addressing these limitations and extending the proposed approach. A more diverse ensemble can be formed by including recent architectures for example ConvNeXt, DenseNet, and Vision Transformers (ViT) to get a wider range of features. DenseNet enhances feature reuse by dense connectivity between layers. Looking into Bayesian or uncertainty-aware ensemble methods might make model calibration better and give confidence estimates alongside predictions. Also, adding Grad-CAM++ or Layer-wise Relevance Propagation (LRP) might help make model explanations easier to understand and measure.

Declarations

Funding. This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Competing interests. The author declares no competing interests.

Ethics approval and consent to participate. This study used the publicly available, de-identified HAM10000 dataset. No new data were collected from human participants, so ethics approval and informed consent were not required.

Data availability. The HAM10000 dataset is publicly available from the Harvard Dataverse (https://doi.org/10.7910/DVN/DBW86T).

References

[1] Sulthana, R., Chamola, V., Hussain, Z., Albalwy, F., Hussain, A.: A novel end-to-end deep convolutional neural network based skin lesion classification framework. Expert Systems With Applications **246**, 123056 (2024) https://doi.org/10.1016/j.eswa.2023.123056

[2] Gupta, S., Tsao, H.: Pathology and Epidemiology of Cancer. Springer, Cham (2016)

[3] Jayaseeli, J.D.D., Briskilal, J., Fancy, C., Vaitheeshwaran, V., Patibandla, R.S.M.L., Syed, K., Swain, A.K.: An intelligent framework for skin cancer detection and classification using fusion of squeeze-excitation densenet with meta-heuristic driven ensemble deep learning models. Scientific Reports **15**, 7425 (2025) https://doi.org/10.1038/s41598-025-92293-1

[4] Ray, A., Gupta, A., Al, A.: Skin lesion classification with deep convolutional neural network: Process development and validation. JMIR Dermatology **3**(1), 18438 (2020) https://doi.org/10.2196/18438

[5] Katsch, F., Rinner, C., Tschandl, P.: Comparison of convolutional neural network architectures for robustness against common artefacts in dermatoscopic images. Dermatology Practical & Conceptual (2022) https://doi.org/10.5826/dpc.1203a126

[6] He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778 (2016). https://doi.org/10.1109/CVPR.2016.90

[7] Minarno, A.E., Lusianti, A., Azhar, Y., Wibowo, H.: Classification of skin cancer images using convolutional neural network with resnet50 pre-trained model. JOIV: International Journal on Informatics Visualization **8**(4), 2013 (2024) https://doi.org/10.62527/joiv.8.4.2156

[8] Aboulmira, A., Hrimech, H., Lachgar, M., Hanine, M., Garcia, C.O., Mezquita, G.M., Ashraf, I.: Hybrid model with wavelet decomposition and efficientnet for accurate skin cancer classification. Journal of Cancer **16**(2), 506–520 (2025) https://doi.org/10.7150/jca.101574

[9] Sajid, M., Khan, A.H., Malik, T.S., Bilal, A., Ahmad, Z., Sarwar, R.: Enhancing

melanoma diagnostic: Harnessing the synergy of ai and cnns for groundbreaking advances in early melanoma detection and treatment strategies. International Journal of Imaging Systems and Technology **35**(1) (2024) https://doi.org/10.1002/ima.70016

[10] Khan, M.A., Sharif, M., Akram, T., Damaševičius, R., Maskeliūnas, R.: Skin lesion segmentation and multiclass classification using deep learning features and improved moth flame optimization. Diagnostics **11**(5), 811 (2021) https://doi.org/10.3390/diagnostics11050811

[11] Efat, A.H., Hasan, S.M.M., Uddin, M.P., Mamun, M.A.: A multi-level ensemble approach for skin lesion classification using customized transfer learning with triple attention. PLOS ONE **19**(10), 0309430 (2024) https://doi.org/10.1371/journal.pone.0309430

[12] Bhargavi, M., Renugadevi, R., Sivabalan, S., Phani, P., Ganesh, J., Bhanu, K.: Ensemble learning for skin lesion classification: A robust approach for improved diagnostic accuracy (elslc). In: 2023 3rd International Conference on Innovative Mechanisms for Industry Applications (ICIMIA), pp. 390–395. IEEE, Bengaluru, India (2023). https://doi.org/10.1109/icimia60377.2023.10425888 . https://doi.org/10.1109/icimia60377.2023.10425888

[13] Wibowo, A., Hartanto, C.A., Wirawan, P.W.: Android skin cancer detection and classification based on mobilenet v2 model. International Journal of Advances in Intelligent Informatics **6**(2), 135–148 (2020) https://doi.org/10.26555/ijain.v6i2.492

[14] Tan, M., Le, Q.: Efficientnet: Rethinking model scaling for convolutional neural networks. In: International Conference on Machine Learning, pp. 6105–6114 (2019). PMLR. http://proceedings.mlr.press/v97/tan19a.html

[15] Szegedy, C., Ioffe, S., Vanhoucke, V., Alemi, A.: Inception-v4, inception-resnet and the impact of residual connections on learning. Proceedings of the AAAI Conference on Artificial Intelligence **31**(1) (2017) https://doi.org/10.1609/aaai.v31i1.11231

[16] Tschandl, P., Rosendahl, C., Kittler, H.: The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. Scientific Data **5**(1) (2018) https://doi.org/10.1038/sdata.2018.161

[17] Sherazi, S.W.A., Bae, J.-W., Lee, J.Y.: A soft voting ensemble classifier for early prediction and diagnosis of occurrences of major adverse cardiovascular events for stemi and nstemi during 2-year follow-up in patients with acute coronary syndrome. PLOS ONE **16**(6), 0249338 (2021) https://doi.org/10.1371/journal.pone.0249338

[18] Li, Y., Luo, Y.: Performance-weighted-voting model: An ensemble machine

learning method for cancer type classification using whole-exome sequencing mutation. Quantitative Biology **8**(4), 347–358 (2020) https://doi.org/10.1007/s40484-020-0226-1

[19] Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: 2017 IEEE International Conference on Computer Vision (ICCV), pp. 618–626. IEEE, Venice, Italy (2017). https://doi.org/10.1109/iccv.2017.74 . https://doi.org/10.1109/iccv.2017.74

[20] Lin, T.-Y., Goyal, P., Girshick, R., He, K., Dollar, P.: Focal loss for dense object detection. In: 2017 IEEE International Conference on Computer Vision (ICCV), pp. 2999–3007. IEEE, Venice, Italy (2017). https://doi.org/10.1109/iccv.2017.324 . https://doi.org/10.1109/iccv.2017.324

[21] Codella, N., Rotemberg, V., Tschandl, P., Celebi, M.E., Dusza, S., Gutman, D., Helba, B., Kalloo, A., Liopyris, K., Marchetti, M., et al.: Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic). arXiv preprint arXiv:1902.03368 (2019)

[22] Cassidy, B., Kendrick, C., Brodzicki, A., Jaworek-Korjakowska, J., Yap, M.H.: Analysis of the isic image datasets: Usage, benchmarks and recommendations. Medical Image Analysis **75**, 102305 (2022) https://doi.org/10.1016/j.media.2021.102305

[23] Goceri, E.: Medical image data augmentation: techniques, comparisons and interpretations. Artificial Intelligence Review **56**(11), 12561–12605 (2023) https://doi.org/10.1007/s10462-023-10453-z

[24] Garcea, F., Serra, A., Lamberti, F., Morra, L.: Data augmentation for medical imaging: A systematic literature review. Computers in Biology and Medicine **152**, 106391 (2023) https://doi.org/10.1016/j.compbiomed.2022.106391

[25] Sahajuddin, Shanto, A.-A.A., Maruf, A.A., Khanam, F.: A novel weighted ensemble strategy using transfer learning for enhanced cervical cancer classification. In: Proceedings of the 3rd International Conference on Computing Advancements (ICCA 2024), pp. 63–70. ACM, New York, NY, USA (2024). https://doi.org/10.1145/3723178.3723187 . https://dl.acm.org/doi/10.1145/3723178.3723187

[26] Singh, J., Beeche, C., Shi, Z., Beale, O., Rosin, B., Leader, J., Pu, J.: Batch-balanced focal loss: a hybrid solution to class imbalance in deep learning. Journal of Medical Imaging **10**(05), 051809 (2023) https://doi.org/10.1117/1.jmi.10.5.051809

[27] Alsaidi, M., Jan, M.T., Altaher, A., Zhuang, H., Zhu, X.: Tackling the class imbalanced dermoscopic image classification using data augmentation and gan. Multimedia Tools and Applications **83**(16), 49121–49147 (2023) https://doi.org/

10.1007/s11042-023-17067-1

[28] V, U., Shivaram, J.M., Guggari, S., Okoye, K.: Ensemble method of pre-trained models for classification of skin lesion images. Applied Sciences **15**(24), 13083 (2025) https://doi.org/10.3390/app152413083 . Published online 12 December 2025, CC BY 4.0 License

[29] Khan, M.A., Rastogi, D., Johri, P., Al-Taani, A., Baghela, V.S., Kumud: Hybrid deep cnn model for multi-class classification of skin lesion. Neural Computing and Applications **37**(23), 19479–19499 (2025) https://doi.org/10.1007/s00521-025-11409-w . Published online 9 July 2025, CC BY License

[30] Akter, M.S., Shahriar, H., Sneha, S., Cuzzocrea, A.: Multi-class skin cancer classification architecture based on deep convolutional neural network. In: 2022 IEEE International Conference on Big Data (Big Data), pp. 5404–5413. IEEE, Osaka, Japan (2022). https://doi.org/10.1109/bigdata55660.2022.10020302 . Published 17 December 2022. https://doi.org/10.1109/bigdata55660.2022.10020302