Proceedings of the 5th International Conference on Advanced Science & Engineering ICOASE 2025, University of Zakho and Duhok Polytechnic University, 24th – 25th September 2025

(Special issue for Passer Journal of Basic and Applied Sciences)
Paper No. 11

Towards Transparent Decisions: CNN Ensemble with XAI-Driven Interpretations

Ahwaz Darweesh Hayder [1], Jwan Najeeb Saeed [2]

[1] Information Technology Management Department, Technical College of Administration, Duhok Polytechnic University, Duhok 42001, Kurdistan Region, Iraq.
[2] Artificial Intelligence Department, Technical College of Duhok, Duhok Polytechnic University, Duhok 42001, Kurdistan Region, Iraq.
Email: ahwaz.hayder@dpu.edu.krd and jwan.najeeb@dpu.edu.krd

Abstract
Skin cancer remains a global health threat with increasing incidence rates. Accurate and early classification of skin lesions into benign or malignant types is essential for timely treatment and prevention of severe outcomes. In this paper, we present a comprehensive deep learning-based framework that leverages three benchmark datasets—PH2, ISIC (Benign vs Malignant), and HAM10000—using transfer learning and ensemble techniques. Pre-trained models including VGG16, ResNet50, and EfficientNetB4 were fine-tuned on each dataset, and majority voting was employed to combine predictions. The Gradient-weighted Class Activation Mapping (Grad-CAM) was also used to improve visual explainability. The findings demonstrate a notable increase in classification accuracy, surpassing current techniques and reaching over 98% accuracy on certain datasets. This study highlights the impact of hybrid architectures and explainable AI in advancing the state of skin cancer diagnosis systems.

Keywords: XAI, Skin lesion classification, ensemble learning, majority voting.

1 Introduction

Skin cancer is still one of the most prevalent and quickly spreading types of cancer in the world, and melanoma has a particularly high risk of death if it is not detected early. Ozone depletion, genetic susceptibility, and increased exposure to sunlight have contributed to a significant increase in skin lesions, according to recent medical studies. [1]. Traditional diagnostic procedures rely largely on clinical experience and visual dermatological assessment, which, although important, is sometimes subject to human error and subjective bias. The remarkable capacity of convolutional neural networks (CNNs) to identify intricate skin images has improved the detection of both benign and malignant diseases. [2]. The increasing need for precise, efficient, and automated diagnostic tools led to the development of deep learning (DL) techniques, a new method for evaluating medical images. Nonetheless, the effectiveness of individual DL models can differ, and extending their outcomes to additional datasets may be challenging. Additionally, despite the significant predictive capabilities of these models, they are frequently perceived as "black boxes" and demonstrate a lack of clarity in how decisions are made. This research significantly tackles the lack of interpretability in image classification and. Consequently, the key contributions of this work are:

- We propose an ensemble classification approach that combines three well-established CNNs—VGG16, ResNet50, and EfficientNetB4—using a majority voting mechanism. This ensemble approach increases overall robustness and makes use of the advantages of many architectures to produce predictions that are more accurate.
- Using Grad-CAM, which produces class-discriminative visual explanations, this research integrates Explainable AI (XAI) approaches to improve interpretability and transparency. By focusing on relevant visual areas, the set helps users and practitioners understand why they made a particular choice.
- The proposed method is rigorously evaluated on three prominent skin lesion datasets—PH2, ISIC, and the combined HAM10000. A thorough performance analysis across many image sources and skin cancer subtypes is made possible by these varied datasets.

Unlike prior ensemble-based approaches, this work focuses on a systematic and reproducible evaluation of a lightweight majority-voting ensemble comprising heterogeneous CNN architectures. Specifically, VGG16 was selected for its ability to capture fine-grained texture features, ResNet50 for learning deep residual representations, and EfficientNetB4 for scalable multi-resolution analysis. By integrating these complementary architectures, the ensemble enhances classification performance across diverse skin lesion types and is rigorously evaluated across multiple datasets and both binary and multi-class classification settings—a level of comprehensive assessment not previously reported in the literature. The remainder of this paper is organized as follows: Section 2 presents the related work. Section 3 describes the methodology. Section 4 presents and analyzes the experimental results and a discussion. Section 5 concludes the work.

2 Related Work

Understanding how decisions are made by AI models has become just as important as achieving high accuracy. Over the years, many studies have tackled this challenge, offering different ways to improve both performance and transparency. [3] introduced a deep ensemble classification framework for skin lesion analysis that aggregates predictions from several deep CNN architectures and evaluated on benchmark datasets. [4] proposed a novel transfer learning-based implementing EfficientNetB0 backbone for multimodal skin lesion analysis by combining both image and textual metadata to improve diagnostic performance. However, it suffers from interpretability challenges typical of complex deep learning architectures. Furthermore the performance can be affected by image quality variations in real-world settings. [5] developed a hierarchical binary classification system using transfer learning and Random Forest algorithms. The first level of the model performs binary classification between cancerous and non-cancerous lesions, while the second level identifies the subtype. It is worth mentioning that although Their two-tier pipeline offered a clear decision-making structure, the need for computational efficiency and limited generalizability across diverse data remains an issue. . [6] performed a comparative analysis of various DL architectures including InceptionV3, DenseNet, and ResNet for skin cancer detection. The DenseNet121 model showed the best performance, achieving an accuracy of 95% across eight different lesion classes. [7] integrated localized, contextual, and hierarchical features using DenseNet121and achieved an accuracy of 93.8% on multi-class data. However, the approach is computationally expensive due to segmentation and classification being processed separately, leading to longer training and inference times. [8] utilized the VGG19 architecture on the ISIC2018 dataset for both segmentation and classification of skin lesions, achieving a classification accuracy of 94.6% across three primary lesion categories.

Merging various technologies can result in more innovative and flexible solutions for difficult problems. To increase the diagnostic accuracy of skin cancer categorization, [9] proposed a hybrid deep learning architecture that incorporates recurrent neural networks (RNNs) with CNNs. Hybrid strategies employ multiple methods to enhance outcomes. While the fundamental neural network (RNN) is capable of recognizing sequential connections in image captions, the convolutional neural network (CNN) can grasp the intricate spatial features of skin images. [10] Research has been done on the use of radiographic analysis in combination with radiography to understand skin lesion patterns. Despite strong results, the study relies on the integration of metadata and radiomic features. The absence of these components in real clinical settings could impact performance. The ablation study also revealed performance drops when removing metadata, highlighting the system's dependency on complete data. In relation to hibernation [11] The researchers presented the ViT-Grad CAM model, a hybrid model that combines the interpretability of vision transducers (ViT) with Grad-CAM. . By enhancing classification performance and offering lucid heat map visuals, this technology assists dermatologists in validating AI-based predictions. [12] Based on multispectral imaging of cutaneous autofluorescence, the researchers proposed a multispectral classification method. To generate diagnostic metrics, they utilized convolutional neural networks, careful image preprocessing, and manually designed radiological imaging features. This integrated approach achieved a 96.1% accuracy rate. This technology offers a unique diagnostic perspective on lesion formation by capturing fluorescence decay signals at different wavelengths. When combined with photoacoustic fluorescence (PAU), the model showed promising outcomes in the early identification of melanomas and precise localization of cancerous cells. This biologically based method enhances the quality of inputs utilized in ultimate classification models. By optimizing features, machine learning models concentrate on the most crucial information. The ant colony optimization (ACO) method has been utilized alongside DL by [13] to enhance skin lesion segmentation. The lesion border matching will be enhanced as a result, particularly when there are abnormalities and low contrast. [14] tested on the ISIC dataset using a CNN and the cat swarm optimization algorithm (CSO), producing a classification accuracy of 91.7%.

3 Theoretical Background

3. 1 The Pretrained Networks

Three pre-trained CNNs —VGG16, EfficientNetB4, and ResNet50—that are frequently utilized for feature extraction and skin lesion picture classification are used in this study. Each of these models offers distinct architectural improvements that enhance learning ability and generalization effectiveness.

A. **VGG16:** is a well-known deep convolutional neural network architecture created by the University of Oxford's Visual Geometry Group (VGG). VGG16 is well-known for its straightforward design and reliability, attributed to its systematic application of 3×3 convolution filters with a stride of 1, accompanied by 2×2 max pooling layers to reduce spatial dimensions. It has 16 weighted layers, 13 of which are convolutional and 3 of which are completely linked. Employing a cascade architecture, this approach collects hierarchical image data by progressively enhancing the network depth. VGG16 concludes with a softmax classification layer and two fully connected layers as shown in Figure 1 below. [15].

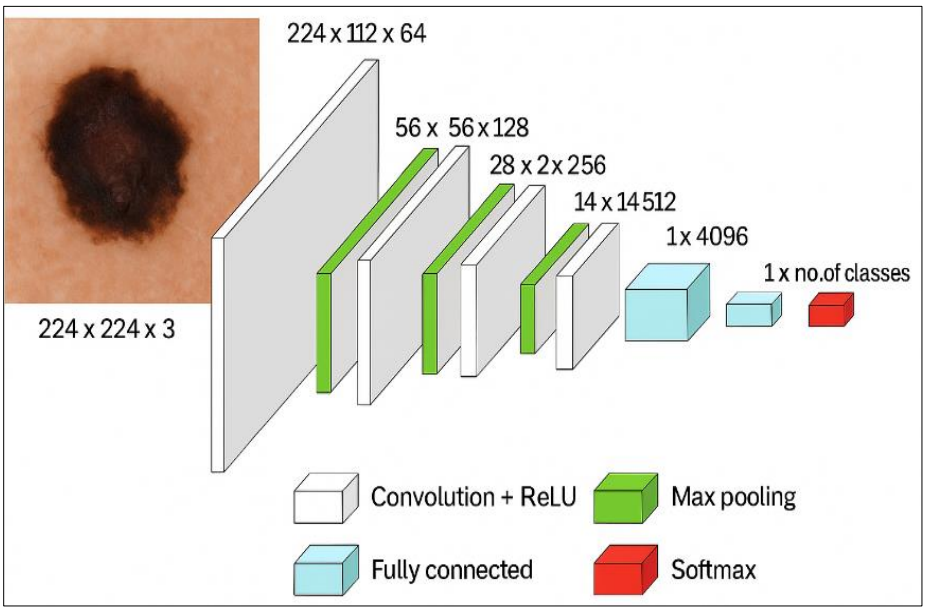

Figure 1: VGG16 architecture.

B. **ResNet50**: is a part of the Residual Network (ResNet) family developed by [16], [17] introduced a residual connectionConvolutional connections, batch normalization, and identity shortcuts that avoid one or more layers are among the 50 layers that make up ResNet50. By permitting gradients to flow directly through identity connections across layers, this method resolves the vanishing gradient issue and makes it possible to build much deeper networks. It employs a modular design based on residual blocks, each of which adds the input of the block directly to its output as shown in Figure 2 below.

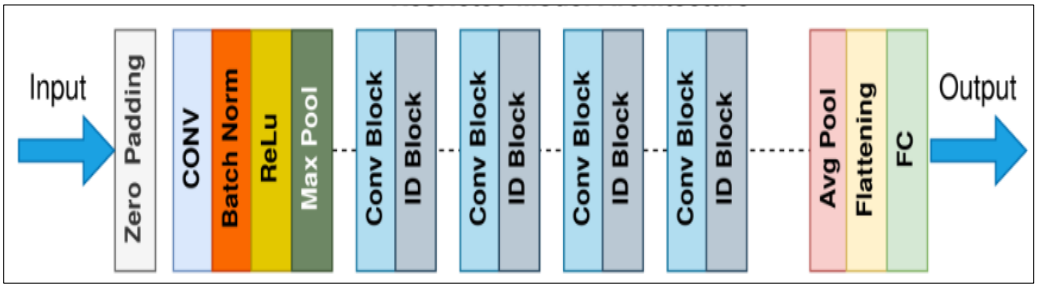

Figure 2: The architecture of ResNet50.

C. **EfficientNetB4:** It is one of the parts of the EfficientNetB4 family developed by [18], which measures depth, width, and resolution using a composite measurement approach. Unlike traditional CNNs that scale only one dimension, EfficientNetB4 applies a balanced scaling strategy to all three dimensions, improving both performance and computational efficiency. EfficientNetB4 comprises multiple mobile inverted bottleneck (MBConv) blocks, swish activation functions, and squeeze-and-excitation optimization mechanisms. It is deeper and wider than its predecessors (B0–B3), providing enhanced accuracy without a substantial increase in computation as shown in Figure 3 below.[19], [20].

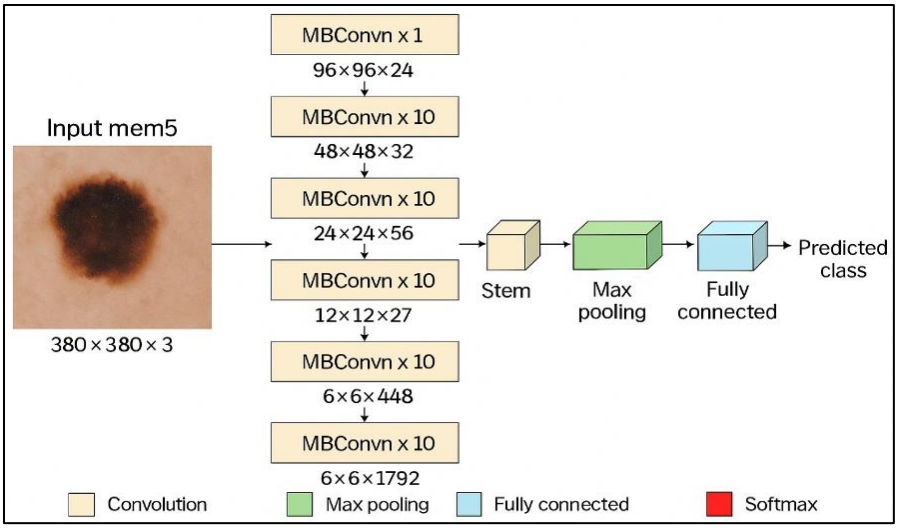

Figure 3: The architecture of the EfficientNetB4.

3. 2 Ensemble Learning

To improve classification accuracy and reduce the likelihood of errors, we employed an ensemble learning approach. Ensemble learning combines the predictions of multiple models to enhance overall performance. In this case, we used a majority voting strategy, where the final classification decision is based on the majority vote from the three models. Every model's prediction was regarded as a "vote," and the final prediction was chosen from the class with the most votesThe ensemble learning approach leverages the strengths of different models and helps mitigate the impact of individual model biases, resulting in a more robust skin cancer classification system.[21].

3. 3 Grad-CAM for Explainability

Grad-CAM (Gradient-weighted Class Activation Mapping) was used to interpret the models' decision-making process. It shows the key regions of the input images that influenced the model's predictions. The Grad-CAM technique is used in Explainable AI (XAI) to give visual explanations of CNN predictions as depicted in Figure 4. The input is first passed through the CNN, and the feature maps Ak from the last convolutional layer, along with the raw class scores (logits) before softmax, are extracted. A target class c typically the one associated with the highest prediction score is then selected, and the gradients of the class scorey^c with respect to each feature map are computed, expressed as $\frac{\partial y^c}{\partial R^k_{ij}}$

The contribution of each feature map to the prediction of class ccc is thereby captured. To summarize the importance of each feature map, global average pooling is applied across the spatial dimensions, resulting in the weights α_k^l defined as

$$\alpha_k^l = \frac{1}{Z}\sum_i \sum_j \frac{\partial y^l}{\partial R^k_{ij}} \qquad (1)$$

, where Z represents the total number of spatial locations. The Grad-CAM heatmap is subsequently generated by weighting the feature maps by their corresponding α_k^l values and applying a ReLU activation:

$$L^c{}_{Grad\_CAM} = \text{ReLU}\left(\sum_k \alpha_k^l A^k\right) \qquad (2)$$

Finally, the heatmap $L^{c}_{Grad\_CAM}$ highlight the regions that most significantly contribute to the model's decision, thus enhancing interpretability within the XAI framework.

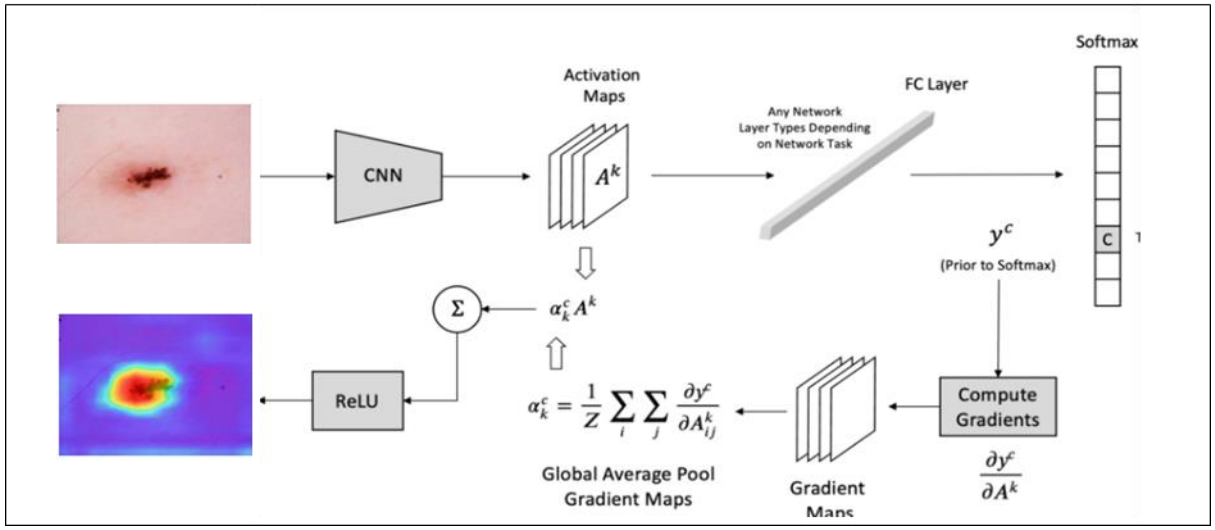

Figure 4: The Grad-CAM general framework.

3. 4 Evaluation Metrics

To evaluate the performance of the individual models and the ensemble approach, we used several standard classification metrics, these include Accuracy, Precision, Recall (Sensitivity), F1-Score, and the Confusion Matrix. The metrics are defined as follows:

- **Precision:** Accuracy assesses a model's ability to predict positive outcomes. It is essential in cases where false positives could lead to severe therapeutic consequences. Formula:

$$Precision = \frac{(TP)}{TP + FP} \qquad (3)$$

Shows the percentage of events rated as positive that were positive.

- **Recall (Sensitivity)**

$$Recall(Sensitivity) = \frac{TP}{TP + FN} \qquad (4)$$

calculates the proportion of real positive cases that the model accurately detected.

- **Specificity**: This model calculates the proportion of true negatives (such as benign lesions) that the model accurately detects. This process is essential for assessing how well the model reduces false alarms, which can lead to unnecessary interventions.

$$Specificity = \frac{TN}{TN + FP} \qquad (5)$$

- **Accuracy:** This model measures the proportion of samples that were correctly predicted (either positive or negative) out of the total predictions. It gives an overall idea of the model's performance across all categories. The formula is:

$$Accuracy = \frac{TP + T\,N}{TP + TN + FP + FN} \qquad (6)$$

Evaluates how accurate the model's predictions are overall.

- **F1-Score:** The harmonic meaning of recision and Sensitivity is used to produce a single statistic that balances as:

$$\text{F1 } Score = 2 \times \frac{Precision \times Sensitivity}{Precision + Sensitivity} \quad (7)$$

4 Methodology

4. 1 Datasets

In this study, three well-known public datasets have been utilized and described as follows

1. **PH2 Dataset**: It contains 200 dermoscopic images classified into three categories: moles, skin cancer, and seborrheic keratoses. Expert dermatologists selected the best images for this collection.

2. **ISIC Dataset**: A wide variety of 3,297 images skin lesion images are available from the International Skin Imaging Collaboration (ISIC) database for binary and multiclass classification applications. The wide variety of image formats and lesion classifications in the ISIC database makes it a popular choice for evaluating the effectiveness of skin cancer detection systems.

3. **HAM10000 Dataset**: It has 10,000 images, and it covers a wide range of skin lesions (7 classes), such as benign lesions, melanomas, basal cell carcinomas, and squamous cell carcinomas. This dataset is a powerful tool for evaluating and training deep learning models.

4. 2 Preprocessing

It is crucial to preprocess the data samples to ensure their consistent quality. This works includes the following preprocessing steps:

- **Data Sampling**: To ensure balanced and robust training across all experiments, a structured data sampling strategy was applied to PH2, ISIC2019-2020, and HAM10000 datasets. The sampling process involved dividing each dataset into training, validation, and testing subsets while preserving the class distribution to avoid bias—especially important given the natural imbalance in skin lesion categories. Splitting the data samples into 80% to serve as training set, 10% per each validation and testing sets individually.

- **Resizing**: All images from PH2, ISIC2019-2020, and HAM10000 datasets were resized accordingly to have the input layer size a spatial resolution of 224 × 224 pixels for both VGG16 and ResNet50 and spatial resolution of 380×380 to match the required input dimensions of the utilized models. Since each model expects a fixed input size based on ImageNet pre-training, resizing ensures consistent feature extraction and compatibility across networks. This step standardizes the spatial scale of dermoscopic lesions, preserves essential visual patterns, and allows the three architectures to process the images efficiently without shape mismatches during training and inference.

- **Normalization:** The pixel values were standardized. Pixel values are frequently scaled in this way to improve their suitability for training. Consequently, In all experiments, pixel intensities of dermoscopic images from the PH2, ISIC, and HAM10000 datasets were normalized prior training . The original RGB images $\boldsymbol{I}$ have pixel values $\boldsymbol{I}(\boldsymbol{i},\boldsymbol{j},\boldsymbol{c})$in the range $[\mathbf{0},\mathbf{255}]$, where $\boldsymbol{i}$and $\boldsymbol{j}$denote the spatial coordinates and $\boldsymbol{c} \in \{\boldsymbol{R},\boldsymbol{G},\boldsymbol{B}\}$is the color channel. Using the rescale = 1. /255 parameter in the Keras ImageDataGenerator, each pixel value was linearly scaled to the range $[\mathbf{0},\mathbf{1}]$ according to:

$$I_{\text{norm}}(i,j,c) = \frac{I(i,j,c)}{255} \quad (8)$$

This uniform normalization was applied to the training, validation, and test splits of all three datasets, ensuring a consistent intensity range across inputs. Such scaling improves numerical stability, accelerates convergence, and better aligns the input distributions with the ImageNet pre-training of VGG16, ResNet50, and EfficientNetB4, leading to more reliable optimization and performance.

- **Data Augmentation**: To enhance the diversity of the training data and mitigate overfitting in deep learning models, a comprehensive data augmentation pipeline was applied to the training subsets of the PH2, and HAM10000 datasets. The augmentation transformations included random rotations (up to 40°), zooming (up to 0.3), horizontal and vertical flips, and width/height shifts (±20 % of the image dimension). These transformations simulate realistic variations in dermoscopic imagery (e.g., orientation, scale, shift), increase the effective sample size, and improve the generalization capability of the three pre-trained. Thus, data balancing was applied only to the training sets using augmentation techniques to equalize class distributions. This approach enhances model generalization, reduces class-bias effects, and ensures fair learning across minority and majority lesion classes, while keeping validation and test sets untouched to preserve real-world data distributions. The following augmentation were configured in Table 1.

Table 1: The data augmentation configuration

Augmentation Type	Value / Range
Rotation	±30–40 degrees
Width/Height Shift	±20%
Zoom Range	20–30%
Shear Transformation	±20%
Horizontal Flip	Enabled

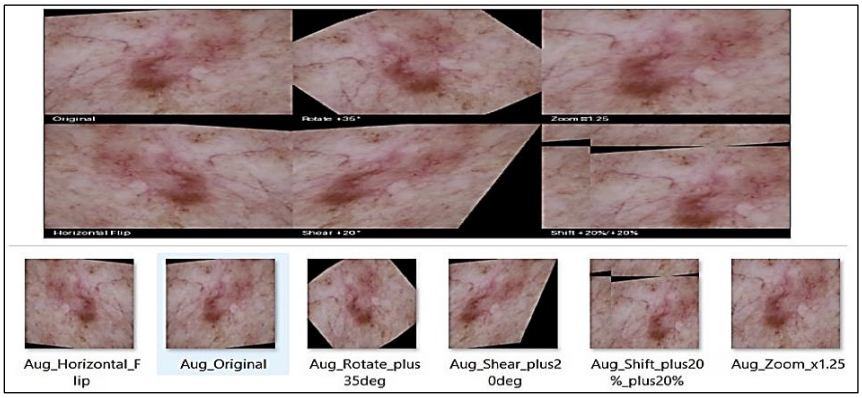

Figure 5. Data Augmentation Samples.

4. 3 The proposed model

This study presents methodology for the classification of skin lesions by fine tuning three deep pretrained networks namely, VGG16, ResNet50, and EfficientNetB4 and fine-tuned them to classify skin lesions into the desired categories. The choice of these models was based on their performance on image classification tasks and their ability to extract features efficiently. The VGG16 offers simplicity and depth, ResNet50 introduces residual connections for improved gradient flow, and EfficientNetB4 balances accuracy and efficiency through compound scaling. This combination was chosen to leverage their individual strengths and enhance overall classification performance. Additionally, preliminary experiments showed that these models consistently outperformed several alternatives on our dataset in terms of both accuracy and generalization. Then the trained models are ensembled. A majority voting ensemble strategy is applied to improve final predictions. The rationale behind the choice of this fusion strategy is due to its simplicity, robustness, and effectiveness in scenarios where base classifiers exhibit comparable performance. In our framework, the individual models were trained to achieve relatively balanced accuracies, minimizing the risk of bias that could arise from other fusion approaches such as weighting schemes. Moreover, to enhance model interpretability, the XAI is utilized through implementing the Grad-CAM technique as shown in Figure 6. In addition, it is worth mentioning that while using three deep models in an ensemble can increase inference time and computational demands, which could be a challenge in environments with limited resources, the significant

improvement in model performance makes the added computational cost worthwhile. This boost in accuracy and reliability is especially valuable for more accurate skin lesion classification.

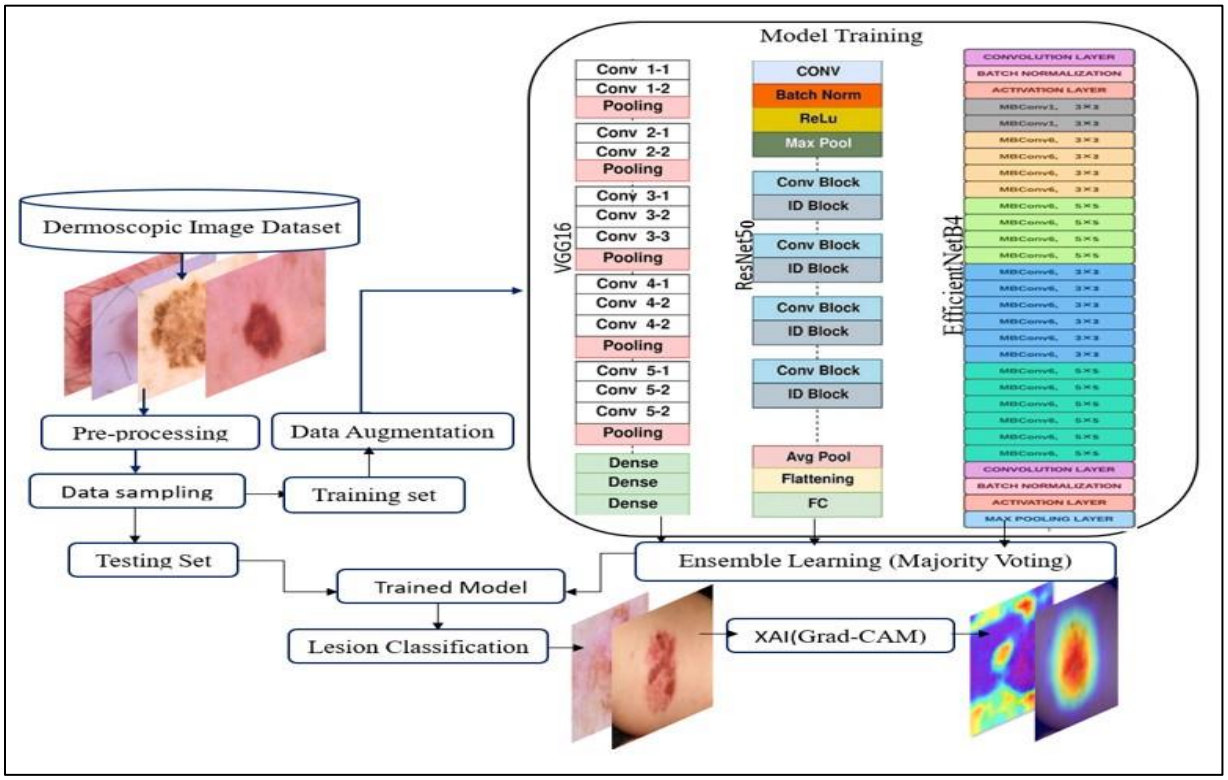

Figure 6: The workflow of the proposed model.

4. 4 Model Training

Each model was individually fine-tuned on the pre-processed datasets, using categorical cross-entropy loss for multi-class classification tasks and binary cross-entropy loss for binary classification. The Adam optimizer was used with an initial learning rate of 0.0001, and a batch size of 64. These hyperparameters were selected after careful experimentation and grid search to balance model training stability and convergence speed. The maximum number of epochs was set to 100, with early stopping (patience of 20 epochs) applied to avoid overfitting, based on performance on the validation set. Additionally, dropout regularization with a rate of 0.5 was employed to further improve generalization. Table 2 shows the setting of the hyperparameters. Once the individual models were trained, their predictions were combined using a majority voting strategy to form the final ensemble output. This ensemble approach was chosen to capture the complementary strengths of different models, reduce variance, and improve overall prediction robustness. The hyperparameter settings were chosen to ensure that each model was trained effectively while maintaining the ability to generalize well to unseen data.

Table 2: The hyperparameters setting

Hyperparameter	**Value**
Optimizer	Adam
Initial Learning Rate	0.0001
Batch Size	64
Epochs	100
Early Stopping	20
Loss Function	Softmax/Cross-Entropy Loss
Dropout	0.5

5 Results and Discussion

To comprehensively evaluate the performance of the proposed model, three benchmark skin lesion datasets PH2, ISIC and HAM10000 were used for training and testing the deep learning models. VGG16, ResNet50, and EfficientNetB4 are among the models that were assessed; each was used separately before being combined by majority vote to increase robustness. Performance was assessed using standard classification metrics such as accuracy, precision, recall, F1-score, and confusion matrix as descried in following subsection.

5. 1 Results on PH2 Dataset

The performance evaluation of the models based on the PH2 dataset is presented in both Table 3and Figure 7. Among the different models, EfficientNetB4 produced the best results, with 96.95% accuracy, 96% precision, 97% recall, and an F1 score of 96.5%. While VGG16 demonstrated strong but somewhat poorer results, ResNet50 also fared well, trailing only EfficientNetB4.The proposed ensemble model performed significantly better than any individual model, achieving 98.33% accuracy, and 98% precision, recall, and F1 score. This consistent and tangible improvement demonstrates the effectiveness of the ensemble approach, which uses majority voting to combine the advantages of different models. In all evaluation benchmarks, the ensemble strategy yielded more stable and balanced performance by leveraging the complementary capabilities of VGG16, ResNet50, and EfficientNetB4 as shown in Figure 1 and Table 3 below.

Table 3: The performance evaluation on PH2 dataset

Model	Accuracy	Precision	Recall	F1-Score
VGG16	0.9583	0.94	0.93	0.935
ResNet50	0.9612	0.95	0.96	0.955
EfficientNetB4	0.9695	0.96	0.97	0.965
Ours	**0.9833**	**0.98**	**0.98**	**0.98**

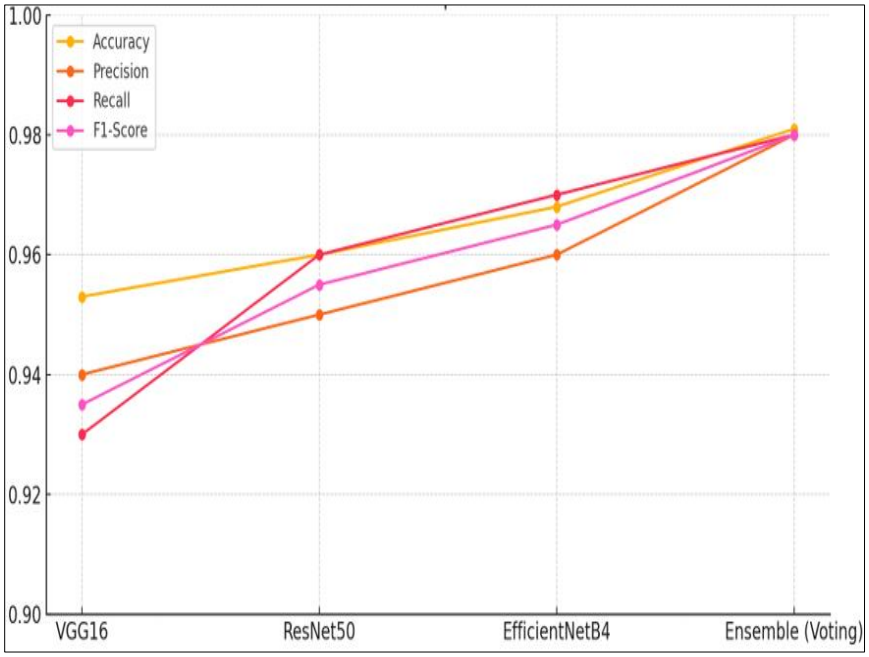

Figure 7: The model performance comparison on PH2

5. 2 Results on ISIC Dataset (Benign vs. Malignant)

Table 4 shows the performance of the proposed methodology on the binary classification task of distinguishing between benign and malignant skin lesions. This demonstrates its excellent ability to accurately recognize both classes

while balancing sensitivity and accuracy. ResNet50 and VGG16 also demonstrated solid performance, achieving accuracy levels of 0.9561and 0.9472, respectively, with consistent values across all evaluation metrics. Meanwhile, the EfficientNetB4 reached an accuracy of 0.9625. The proposed ensemble model earned an astounding accuracy of 0.9736, clearly beating all individual models. This significant improvement highlights the advantage of combining the strengths of multiple architectures through majority voting, leading to a more robust and reliable classifier. The ensemble approach not only enhanced overall predictive performance but also contributed to greater stability in classifying challenging cases, such as borderline lesions that might confuse individual models.

Table 4: The comparative evaluation on ISIC dataset

Model	Accuracy	Precision	Recall	F1-Score
VGG16	0.9472	0.95	0.95	0.95
ResNet50	0.9561	0.96	0.95	0.95
EfficientNetB4	0.9625	0.97	0.97	0.97
Ours	0.9736	0.99	0.99	0.97

5. 3 Results on HAM10000 (multi-Class, 7 Classes)

Table 5 presents a comparative evaluation of the individual models — VGG16, ResNet50, and EfficientNetB4 — alongside the proposed ensemble model, all tested on the HAM10000 dataset. As shown, VGG16 achieved an accuracy of 87.12%, with closely aligned precision, recall, and F1-scores around 84–85%, indicating balanced but modest performance. ResNet50 outperformed VGG16 across all metrics, reaching an accuracy of 88.74% and maintaining high consistency between precision 88.90%, recall 89.10%, and F1-score 89.00%. EfficientNetB4 demonstrated an even stronger performance, achieving an accuracy of 89.62%, along with precision and recall values exceeding 91%, reflecting its capacity to capture more complex patterns within the dataset. However, the best results were obtained by the proposed ensemble model, which combined the predictions of the individual models through majority voting. The ensemble achieved an impressive accuracy of 91.85%, with precision, recall, and F1-scores consistently above 94%.

Table 5: The performance evaluation on HAM10000

Model	Accuracy	Precision	Recall	F1-Score
VGG16	0.8712	0.8460	0.8500	0.8480
ResNet50	0.8874	0.8890	0.8910	0.8900
EfficientNetB4	0.8962	0.9130	0.9160	0.9145
Ours	0.9185	0.9430	0.9450	0.9440

The performance comparison of the models across three datasets with varying complexity and characteristics as illustrated in Table 6 and Figure 8. Regarding the PH2 dataset, being a relatively small and more homogeneous dataset with high-quality dermoscopic images, allowed all models to achieve high accuracy, with EfficientNetB4 slightly outperforming VGG16 and ResNet50. The ensemble model further pushed the performance to 98.33%, benefiting from the consistency within the dataset. Furthermore, The ISIC dataset, which is larger and more diverse in terms of lesion appearance, posed a greater challenge compared to PH2. Here again, EfficientNetB4 showed superior individual performance, and the ensemble approach led to a notable increase in accuracy to 97.36%, highlighting the ensemble's ability to generalize better across more variable data. Finally, the HAM10000 dataset, being the most heterogeneous and complex presented the greatest difficulty for classification. As expected, individual model performances were slightly lower compared to PH2 and ISIC. Nevertheless, the ensemble strategy still resulted in the highest accuracy of 91.85%, confirming that combining multiple models helps to mitigate the challenges posed by high data variability and noise.

Table 6: Model performance comparison across datasets

Dataset	VGG16	ResNet50	EfficientNetB4	Ensemble
PH2	95.83	96.12	96.95	98.33
ISIC	94.72	95.61	96.25	97.36
HAM1000	87.12	88.74	89.62	91.85

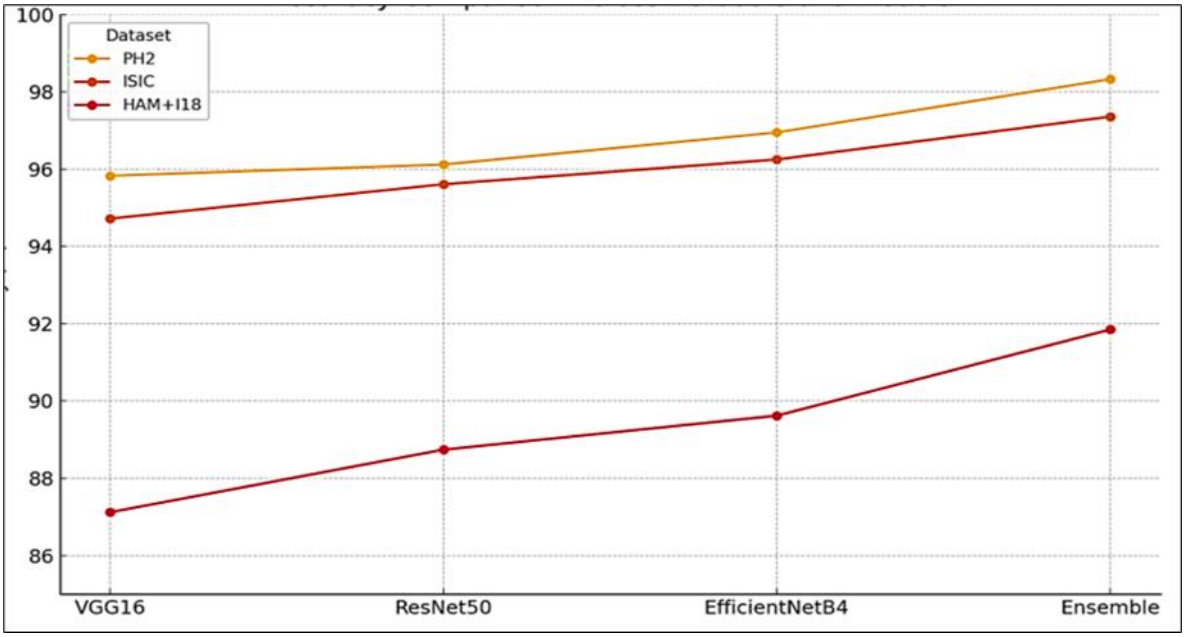

Figure 8: Accuracy comparison across models and datasets.

5. 4 Explainability via Grad-CAM

Dermoscopic images are critical because they contain rich visual information, such as color distribution, texture, and structural features, which are essential for accurate classification. The diversity of the lesions illustrates the challenge of skin lesion classification, where visual similarities between different types can confuse traditional diagnostic methods and AI systems alike. This visual explanation allows clinicians and researchers to verify whether the model is focusing on the correct lesion region rather than unrelated image artifacts, thus increasing trust in the AI system. The Grad-CAM results revealed that the proposed model was able to correctly identify the critical features of skin lesions as depicted in Figure 9 Grad-CAM visual explanations for representative samples from different datasets and lesion types. ISIC2019-2020 dataset, PH2 dataset, and HAM10000 dataset. The first row shows the original dermoscopic images, while the second row presents the corresponding Grad-CAM heatmaps generated by the proposed deep learning model. It demonstrates well-focused regions of activation, with warm areas (red, yellow) showing high levels of importance. In contrast, regions that are not relevant to the classification (such as the background) are mostly colored in cooler tones (blue, green), suggesting lower relevance to the prediction. Consequently, the highlighted regions indicate the areas that contributed most to the model's decision. This visualization is particularly important in the medical field, as it enhances transparency and allows clinicians to understand the reasoning behind the model's predictions. For Benign category, the Grad-CAM focuses on the symmetry and uniformity of the lesion, which are distinguishing features of benign moles compared to malignant melanoma. Therefore, the attention maps are compact and centrally focused, indicating that the model relies primarily on lesion shape and smooth boundaries for classification. In contrast , in Melanoma cases, the Grad-CAM maps clearly emphasize lesion borders and internal texture patterns, which are well-known diagnostic characteristics and align with clinical expertise. Meanwhile, the uneven surface texture and irregular shape of the lesion, reflecting the erythematous undertones typical of actinic keratosis (AKIEC). The heatmap highlights the scaly, flat regions, which correspond to the model's focus on abnormal skin texture, consistent with AKIEC, a pre-cancerous condition commonly found in sun-damaged areas. Moreover, focusing on the uneven pigmentation (lighter and darker brown areas). This aligns with the characteristics of benign keratosis-like lesions (BKL), where the model tends to focus on the overall structure and surface irregularities rather than distinct, localized hotspots. Thus, The Grad-CAM results confirm that the model focuses on clinically relevant features within each lesion type, which is crucial for both the trustworthiness and adoption of AI in medical diagnostics. Moreover, these visualizations contribute to the overall explainability of AI systems, making them more acceptable in medical practices by offering insights into how the model arrived at its conclusions.

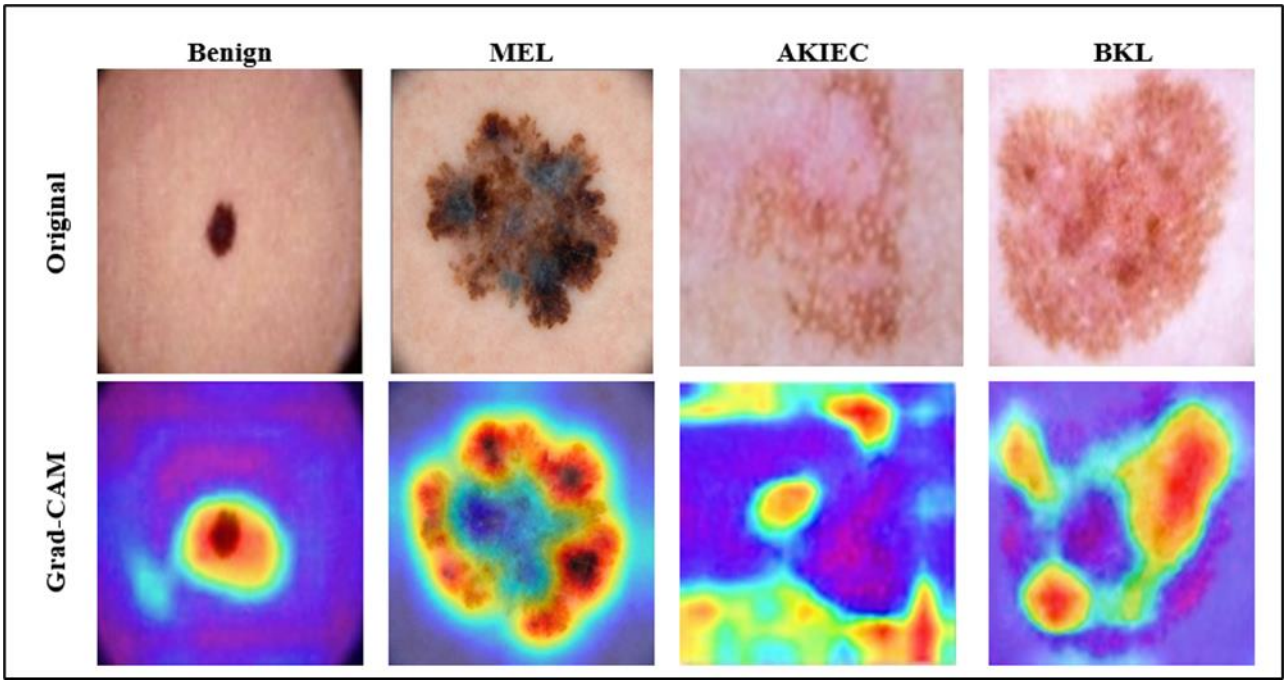

Figure 9: XAI samples based on the proposed ensemble model.

5. 5 Comparison with the state of the art

To situate the proposed ensemble-based model within the existing literature, its performance is compared with several representative approaches for skin lesion classification, as summarized in Table 7 and Figure 10. The selected studies span different model architectures and learning strategies and report results on commonly used benchmark datasets, including PH2, ISIC, and HAM10000. Because these studies address different classification settings—ranging from binary to multi-class tasks—and employ distinct preprocessing and evaluation protocols, the reported performance metrics should be interpreted as indicative rather than directly equivalent. In particular, multi-class classification on datasets such as HAM10000 presents additional challenges due to higher intra-class variability and pronounced class imbalance, which can influence model accuracy. Within this broader context, the proposed model demonstrates competitive performance across all evaluated datasets. On the PH2 dataset with three classes, the model achieves strong accuracy value, indicating effective discrimination under relatively controlled conditions. For binary classification on the ISIC dataset, the proposed approach yields performance comparable to recent ensemble-based methods. On the more complex HAM10000 dataset with seven classes, the model maintains a lower overall accuracy, reflecting balanced predictive behavior in a challenging multi-class setting.

Overall, this comparison suggests that the proposed ensemble model performs consistently across diverse datasets and task formulations. Rather than focusing on absolute performance rankings, the results highlight the model's ability to generalize across varying levels of classification complexity, which is an important consideration for practical deployment in clinical decision-support systems

Table 7: A summary of the compassion of the proposed model and state-of-the-art.

Ref.	Model	Dataset	#class	Accuracy%
[15]	ResNet50 + MDKLoss	HAM10000	7	94.2
[22]	IncepX-Ensemble	ISIC(Binary)	2	95.1
[2]	SkinLesNet (Multi-layer CNN)	PH2	3	96.2
[23]	Ensemble Learning	ISIC (Binary)	2	97.7
[24]	Hybrid Capsule Network	PH2	3	93.6
Ours	The proposed Ensemble Model	PH2	3	**98.33**
Ours	The proposed Ensemble Model	ISIC (Binary)	2	**97.36**
Ours	The proposed Ensemble Model	HAM10000	7	**91.85**

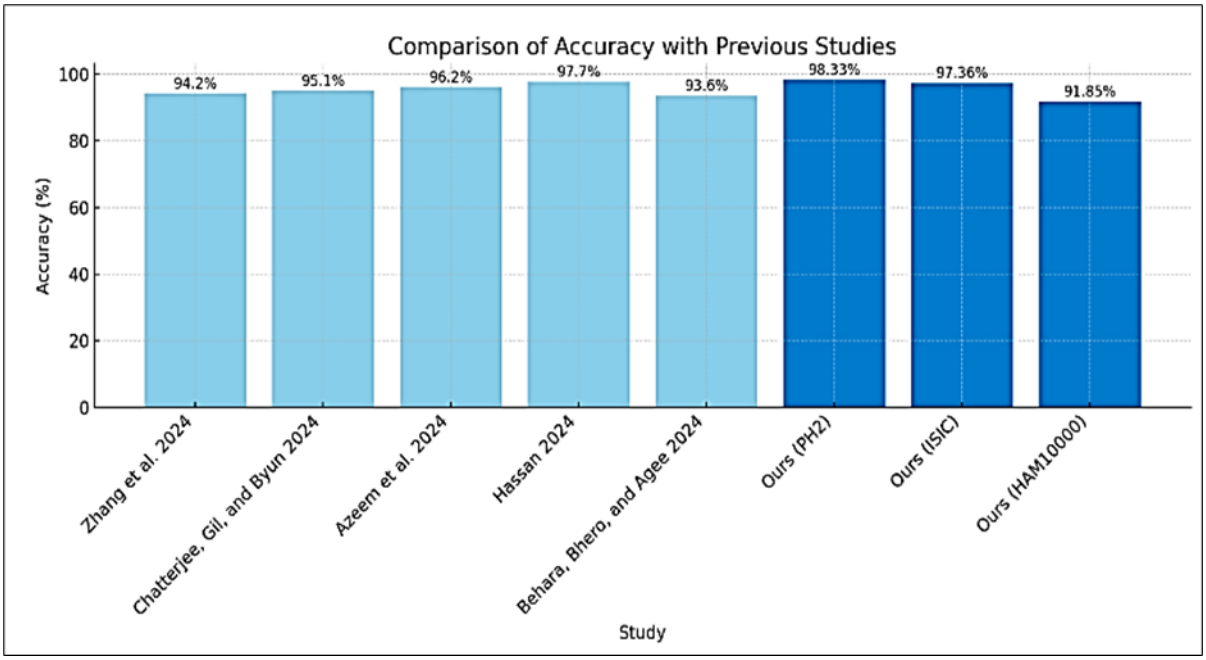

Figure10: Accuracy comparison among the proposed methodology and state of the art.

6 Conclusion

This study proposed an ensemble deep learning framework for improved skin lesion classification, integrating three powerful pre-trained models—VGG16, ResNet50, and EfficientNetB4 and applying majority voting to enhance diagnostic reliability across diverse datasets. Extensive experiments were conducted on PH2, ISIC (Benign vs Malignant), and HAM10000 datasets, and the proposed methodology achieved accuracies of 98.33%, 97.36%, and 91.85% respectively, outperforming many existing approaches reported in recent literature. The inclusion of Grad-CAM visualization played a pivotal role in model interpretability, enabling clear identification of key lesion regions that influenced the model's predictions. This explainability is essential for clinical acceptance and enhances the transparency of AI-based diagnosis. The success of the proposed framework demonstrates the effectiveness of ensemble learning in addressing core challenges in skin lesion classification, such as intra-class similarity, and model generalization. Future work could explore multimodal input fusion, build more efficient models for low-resource settings, and enable real-time adaptation to evolving clinical data.

References

[1] Alsuwaidan, S. *et al.* The Prevalence and Awareness of Regular Follow Up of Patients Treated with Phototherapy for Skin Cancer Surveillance. *Clin. Cosmet. Investig. Dermatol.* **17**, 1739–1745 (2024).

[2] Azeem, M., Kiani, K., Mansouri, T. & Topping, N. SkinLesNet: Classification of Skin Lesions and Detection of Melanoma Cancer Using a Novel Multi-Layer Deep Convolutional Neural Network. *Cancers (Basel).* **16**, 1–15 (2024).

[3] Thwin, S. M. Applied Sciences. *Early Writings on India* 124–134 doi:10.4324/9781315232140-14 (2018).

[4] Remya, S., Anjali, T. & Sugumaran, V. A Novel Transfer Learning Framework for Multimodal Skin Lesion Analysis. *IEEE Access* **12**, 50738–50754 (2024).

[5] Suleiman, T. A., Anyimadu, D. T., Permana, A. D., Ngim, H. A. A. & Scotto di Freca, A. Two-step hierarchical binary classification of cancerous skin lesions using transfer learning and the random forest algorithm. *Vis. Comput. Ind. Biomed. Art* **7**, (2024).

[6] Hussain, M. *et al.* SkinNet-INIO: Multiclass Skin Lesion Localization and Classification Using Fusion-Assisted Deep Neural Networks and Improved Nature-Inspired Optimization Algorithm. *Diagnostics* **13**, (2023).

[7] Ramamurthy, K., Thayumanaswamy, I., Radhakrishnan, M., Won, D. & Lingaswamy, S. Integration of Localized, Contextual, and Hierarchical Features in Deep Learning for Improved Skin Lesion Classification. *Diagnostics* **14**, (2024).

[8] Rani, A., Khalil, A., Shah, Y. A. & Jr, I. Deep Learning-based Skin Lesion Segmentation and Classification. (2024).

[9] Zareen, S. S., Sun, G., Kundi, M., Qadri, S. F. & Qadri, S. Enhancing Skin Cancer Diagnosis with Deep Learning: A Hybrid CNN-RNN Approach. *Comput. Mater. Contin.* **79**, 1497–1519 (2024).
[10] Wang, Z. *et al.* Radiomic and deep learning analysis of dermoscopic images for skin lesion pattern decoding. *Sci. Rep.* **14**, 1–14 (2024).
[11] Shafiq, M. *et al.* A novel Skin lesion prediction and classification technique: ViT-GradCAM. *Ski. Res. Technol.* **30**, 1–12 (2024).
[12] Vasanthakumari, P. *et al.* Pixel-level classification of pigmented skin cancer lesions using multispectral autofluorescence lifetime dermoscopy imaging. *Biomed. Opt. Express* **15**, 4557 (2024).
[13] Sarwar, N., Irshad, A., Naith, Q. H., D Alsufiani, K. & Almalki, F. A. Skin lesion segmentation using deep learning algorithm with ant colony optimization. *BMC Med. Inform. Decis. Mak.* **24**, 265 (2024).
[14] Rajendran, V. A. & Shanmugam, S. Automated Skin Cancer Detection and Classification using Cat Swarm Optimization with a Deep Learning Model. *Eng. Technol. Appl. Sci. Res.* **14**, 12734–12739 (2024).
[15] Zhang, L., Xiao, X., Wen, J. & Li, H. MDKLoss: Medicine domain knowledge loss for skin lesion recognition. *Math. Biosci. Eng.* **21**, 2671–2690 (2024).
[16] He, X., Wang, Y., Zhao, S. & Yao, C. Deep metric attention learning for skin lesion classification in dermoscopy images. *Complex Intell. Syst.* **8**, 1487–1504 (2022).
[17] Zhou, D. *et al.* Distribution characteristics of immune infiltration and lymphovascular invasion in patients with breast cancer skin recurrence. *Cancer Immunol. Immunother.* **73**, (2024).
[18] Tan, M. & Le, Q. V. EfficientNet: Rethinking model scaling for convolutional neural networks. *36th Int. Conf. Mach. Learn. ICML 2019* **2019-June**, 10691–10700 (2019).
[19] Natha, P. & RajaRajeswari, P. Advancing Skin Cancer Prediction Using Ensemble Models. *Computers* **13**, (2024).
[20] Nasim, M. *et al.* Novel paradigm of therapeutic intervention for skin cancer: challenges and opportunities. *Futur. J. Pharm. Sci.* **10**, (2024).
[21] Khan, M. A. *et al.* Multiclass skin lesion classification using deep learning networks optimal information fusion. *Discov. Appl. Sci.* **6**, (2024).
[22] Chatterjee, S., Gil, J. M. & Byun, Y. C. Early Detection of Multiclass Skin Lesions Using Transfer Learning-Based IncepX-Ensemble Model. *IEEE Access* **12**, 113677–113693 (2024).
[23] Hassan, B. Transfer Learning with Imbalanced Datasets for Skin Cancer Detection Abstract. (2024).
[24] Behara, K., Bhero, E. & Agee, J. T. An Improved Skin Lesion Classification Using a Hybrid Approach with Active Contour Snake Model and Lightweight Attention-Guided Capsule Networks. *Diagnostics* **14**, (2024).