VFAST Transactions on Software Engineering
https://vfast.org/journals/index.php/VTSE@ 2026, ISSN(e): 2309-3978, ISSN(p): 2411-6246
Volume 14, Number 1, January-March 2026 pp:60-72

An Ensemble Deep Learning Framework for Automated Multi Class Skin Lesion Classification Using ConvNeXt-Tiny, EfficientNetV2-S, and MobileNetV3

Muhammad Faris[1*], **Falak Khalid**[1], **Maida Shahid**[1], **Zeeshan ul Haque**[1], **Syed Ibad Hasnain**[1], **Mir Farooq Ali**[2]

[1]Department of Biomedical Engineering, Faculty of Engineering, Science and Technology, Hamdard University, Karachi, Pakistan; [2]Dipartimento di Ingegneria dell'Informazione, Università Politecnica delle Marche, Ancona, Italy

Keywords: *Ensemble Deep Learning, EfficientNetV2, MobileNetV3, Dermoscopic Image Analysis, Explainable Artificial Intelligence (XAI), Smooth-Grad*

Journal Info:
Submitted:
December 31, 2025
Accepted:
February 09, 2026
Published:
February 16, 2026

Abstract Skin cancer is one of the fastest rising malignancies in the world, where early and accurate diagnosis plays a decisive role in the survival of patients. Even though the performance of convolutional neural networks (CNNs) in dermoscopic image analysis has been impressive, three main issues that hinder their application in the clinical setting remain: severe imbalance in classes, inter-class visual similarity, overfitting, and lack of interpretability. In this paper, a three-stream ensemble deep learning model that combines ConvNeXt-Tiny, EfficientNetV2-S, and MobileNetV3-Large has been proposed to classify automated multiclass skin lesion on the HAM10000 dataset. Balanced stratified sampling strategy (BalancedDataGen) is used in order to take care of class imbalance without synthetic oversampling in order to have equal contribution of classes in the training. More successful feature diversity and generalization are further improved with the help of adversarial Albumentations-based data augmentation. Each backbone model is fine-tuned separately based on label smoothing, dropout regularization, early stopping, and adaptive learning rate scheduling. The probability-level ensemble averaging is done to give final predictions. The post-hoc explainability of the results is offered to improve clinical interpretability, and ensemble saliency maps created through SmoothGrad allow seeing lesion-specific discriminatory areas consistently. Experimental findings indicate that the proposed ensemble is more accurate, robust and minority-class recognizing than individual models, and it has a test accuracy of 89.72%. The framework provides a good, interpretable and computationally efficient solution to automated dermoscopic diagnosis to fit low-resource clinical settings.

***Correspondence author email address:** m.faris@hamdard.edu.pk
DOI: 10.21015/vtse.v14i1.2311

1 Introduction

Skin cancer is the uncontrolled growth of degenerated cells in the skin and is one of the most prevalent types of cancers in the world. It usually starts in the epidermis, the most superficial skin layer, but the damage to DNA usually caused by ultraviolet (UV) radiations, whether sunlight or in a tanning bed, disrupts normal cell growth and apoptosis, causing malignant transformation and permanent cell division [1]. DNA mutations related to UV induce aberrant signaling pathways, which promote

the emergence of tumors, and a lifetime cumulative exposure is a major risk factor in the development of both non-melanoma and melanoma skin cancer [2].

The most common forms of skin cancer are non-melanoma types (basal cell carcinoma, BCC, and squamous cell carcinoma, or SCC) and melanoma which starts in the pigment-containing melanocytes and is the most fatal among them since it has a high metastatic rate and can spread into other body organs unless it is detected early[3]. BCC and SCC are often situated in sun-exposed areas like the face and neck and are often curable through the use of early surgery intervention, but melanoma has a much smaller proportion of disease but a disproportionate number of deaths [4]. It is estimated that in 2022 alone over 1.5 million new cases of skin cancer are going to be seen globally, including about 330 000 melanomas, tens of thousands of deaths among which are likely to be a result of metastatic disease, highlighting the dramatic impact of this disease on the health of the general population, and the importance of early and correct diagnosis [1, 5].

These aspects of disease and the epidemiology help to pinpoint the clinical need to have effective screening and diagnostic measures, which could be useful in promoting early diagnosis and risk classification. Early and correct identification of skin lesions can contribute greatly to the clinical outcomes, including a decrease in mortality and invasive procedures [6]. Clinicians however, using manual inspection of dermoscopic images, is both time-consuming and highly expertise-dependent, which results in inconsistencies and possible mistake diagnoses. Deep learning-based automated skin lesion classification has, therefore, become a research focus in medical image analysis [7].

Deep learning models have especially convolutional neural networks (CNNs) in recent years dominated automatic classification of skin lesions because it is able to learn hierarchical feature representations directly based on original images[8]. Experiments on the application of CNNs to dermatological images have shown that deep models can outperform expert dermatologists in terms of performance at least and sometimes significantly better, when trained on large image corpora and fine-tuned on dermoscopic datasets [9]. Since then, CNN-based methods have been applied to a variety of architectures, such as DenseNet, Xception, MobileNet, and EfficientNet families to improve the classification performance over seven or more lesion types with different visual appearances [10].

Contemporary developments have additionally discussed the lightweight models and ensemble methods to trade undergoing accuracy and computational efficiency to apply them in resource-contained clinical and mobile conditions. Even with these advances, there are a number of challenges that are still present. First, the imbalance and lack of adequate data prevent generalizable learning, as the rare categories of lesions are underrepresented in popular benchmarks like HAM10000 and ISIC series, which results in bias of the classifier towards majority classes and decreased sensitivity to important rare diseases [11, 12].

Conventional solutions such as oversampling and class-weighted loss functions are prone to synthetic artifacts or learning dynamics. Second, there exist inherent limitations of dealing with inter-class similarity and inter-class variance: lesions that are visually similar (e.g., melanoma and benign keratosis-like lesions) and large variability in the features within the lesion types make the learning of standard CNNs more challenging. To alleviate these problems, hybrid models and attention mechanisms have been suggested, yet they have the drawback of adding more complexity to the model, and consume more training resources [13].

The other emerging area of research is explainability. Although high classification metrics have been reported by many studies, the clinical adoption of deep models has been hindered by their black box characteristics. The methods of Explainable Artificial Intelligence like the Grad-CAM and SmoothGrad offer visual explanations of any model choice, yet neither their interpretability nor clinical usefulness has consistently been investigated [14]. According to recent studies, saliency maps and heatmaps may occasionally be inaccurate to clinical interpretation, which is why more effective XAI tools that are verified with expert annotations are required [15]. Moreover, additions of clinical metadata and multimodal data (e.g., age of a patient, the localization of a lesion) to classification models have reported some success but are not yet widely used in automated systems. The few forming studies have indicated that adding structured

patient-based data with image characteristics in hybrid deep learning models can enhance model resilience and reflect life clinical diagnostic thinking more accurately [16, 17].

Lastly, adaptations across domains and datasets and imaging condition are also a major research challenge. Results indicate a discrepancy between controlled experimental settings and clinical practice as models that are trained on one dataset often poorer in deployment on images in other systems of acquisition or other patient demographics [18, 19].

As a careful analysis of recent literature shows, even though deep learning has made significant progress in the classification of skin lesions, existing paradigms are largely concerned with improving the accuracy of models in one dimension [20]. The majority of the studies lack the discussion of combined issues of data imbalance, inter-class similarity, explainability, and clinical integration in a single framework. Moreover, explainability methods are frequently discussed as after-hoc add-ons, not as part of the methods based on clinical needs. Very little methods have integrated multimodal data and successfully generalized well with divergent datasets and conditioning of imaging.

In these gaps, to achieve the aims of compensating such deficiencies, this paper will put forward a multi-architecture ensemble structure incorporating complementary CNN backbones, balanced sampling of different minority classes, targeted augmentation, and in classifying skin lesions, enable each to plan a robust architecture, minority-class recognition, and interpretability. The proposed work will fill the gap between the role of high-performance experimental models and clinically feasible automated diagnostic systems by ensuring that model design is consistent with practical clinical constraints and diagnostic transparency.

2 Methodology

This section outlines the suggested ensemble deep learning system of automated multiclass of skin lesions. The methodology is designed to address major limitations of dermoscopic image analysis, which are a strong imbalance of classes, high inter-class similarity, scarce data, and the need to be clinically interpretable. The overall pipeline consists of balanced data sampling, targeted data augmentation, multi-architecture transfer learning, probability level ensemble fusion, and post-hoc explainability via ensemble SmoothGrad saliency maps, as shown in Fig. 1.

2.1 Dataset Description

The model being examined was tested using the HAM10000 publicly available benchmark which is widely used in the detection of dermatological images from Harvard Dataverse [21]. The sample is 10,015 high-resolution dermoscopic images collected through a variety of sources and clinical settings. It consists of seven types of diagnosis: Melanoma (MEL), Melanocytic Nevi (NV), Basal Cell Carcinoma (BCC), Actinic Keratoses and Intraepithelial Carcinoma (AKIEC), Benign Keratosis-like Lesions (BKL), Dermatofibroma (DF), and Vascular Lesions (VASC) [21]. There is a sharp imbalance in classes: NV predominates in the distribution of samples, and DF and VASC are significantly underrepresented, as shown in Fig. 2. This is an immense setback to traditional supervised learning systems.

2.2 Data Splitting Strategy

The dataset is stratified by stratified sampling to guarantee dependable evaluation and to maintain the distribution of classes between subsets by dividing the dataset into a training set (70%), a validation set (10%), and a test set (20%). Stratification ensures that all subsets built represent the distribution of diagnostic classes and thus avoids bias when selecting the model and evaluation of performance.

2.3 Image Processing

Before the model training, all of the dermoscopic images will go through the uniform preprocessing pipeline, which ensures uniformity across samples, and adherence to the relevant deep-learning architectures. All pictures are downsampled to a spatial resolution of 224 x 224 pixels, which is the default size of the input needed by ConvNeXt-Tiny, EfficientNetV2-S, and MobileNetV3-Large. This resolution is a sensible trade-off between patients who only characterize lesions clinically (including border irregularities, pigmentation patterns, and textural details) and patients who are simple to compute (in training and inference).

After resizing, pixel values are normalized based on ImageNet statistics and this brings the input distribution

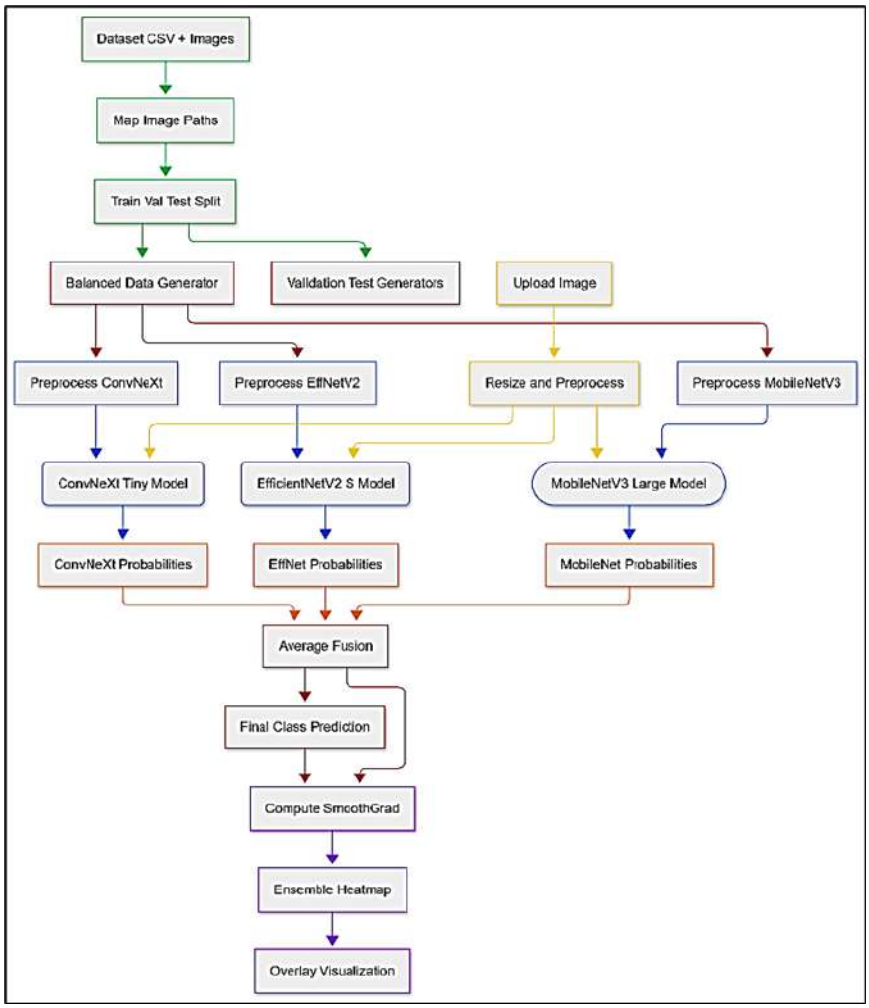

Figure 1. Methodology Workflow

to the pretrained weights. The importance of this step of normalization to transfer learning is to reduce distributional shift between the source domain (ImageNet) and the target domain (dermoscopic images), which in turn stabilizes optimization and speeds up the convergence. As an additional measure to make the architecture compatible, the individual models use their own preprocessing functions, which enables the feature extraction to be in line with the original training setup of the backbone networks.

The preprocessing pipeline itself is implemented as part of the data creating pipeline and run on-the-fly during their training and testing. The design does not incur unnecessary memory overheads and also can be efficiently loaded, transformed, and normalized in batch. The standardization of image resolution, intensity distribution, and preprocessing semantics of all models makes the proposed framework to provide fair comparison of the individual networks and also aids in the ability to fuse probabilities at the probability level.

2.4 Class-Balanced Sampling Using BalancedDataGen

The HAM10000 dataset has a strong imbalance in classes, with the bulk of the samples being melanocytic nevi, and several clinically important groups (including dermatofibroma and vascular lesions) being severely underrepresented. Training deep neural networks with standard random sampling often produces a biased decision boundary, which disadvantages minority classes and thus adversely affects the sensitivity for large rare lesion types. To address this, a custom class-balanced

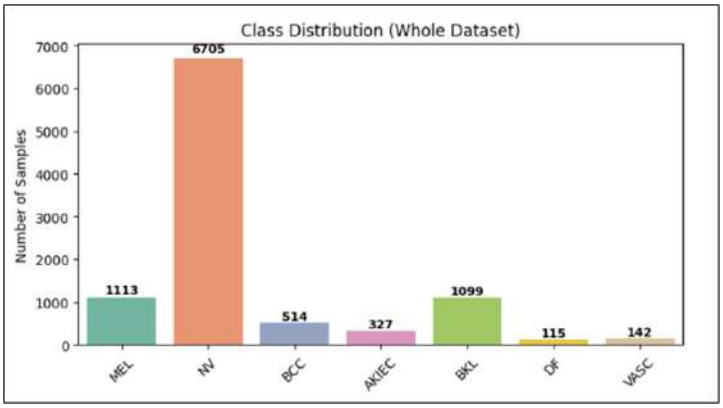

Figure 2. Dataset Class Distribution

sampling scheme, referred to as BalancedDataGen, is applied when training the models (see Figure 3 for the training, testing, and validation split).

BalancedDataGen requires all the classes to contribute equally at the mini-batch level by initially randomly choosing a diagnostic class and then selecting an image at random from the chosen class. This step is repeated until a complete batch is assembled. As a result, each of the seven categories of lesions contributes equally to gradient updates throughout the training regime, having a neutral effect on the loss function. This method is not associated with any synthetic oversampling methods or class-weighted losses; unlike those, it does not inject artificial samples or create notoriously unstable gradients, thus maintaining the natural distribution of dermoscopic features realistically while better representing minority classes (see Figure 4 for a pseudo-code illustration of BalancedDataGen).

Mathematically,

$$c_i \sim \text{Uniform}\big(\{1,\ldots,C\}\big), x_i \sim \text{Uniform}\big(\mathcal{D}_{c_i}\big), i = 1,\ldots,B, \tag{1}$$

where the mini-batch is defined as

$$\mathcal{B} = \{x_1,\ldots,x_B\}, \tag{2}$$

and the class prior distribution satisfies

$$P(y = c) = \frac{1}{C}. \tag{3}$$

The expected number of samples from class c in a batch is therefore:

$$\mathbb{E}[n_c] = \frac{B}{C}. \tag{4}$$

The proposed framework is able to maintain computational efficiency by incorporating BalancedDataGen into the data generator, which avoids the extra preprocessing costs. This sampling approach provides a significant enhancement to the ability of the model to learn discriminative features to all types of lesions and is critical in boosting recall and F1-scores on underrepresented classes.

2.5 Data Augmentation Strategy

In order to increase the generalization and reduce overfitting, a carefully developed data augmentation protocol is applied with the help of Albumentations library. Dermoscopic images have a natural variability due to variations in light levels, variations in imaging equipment, pigmentation of the skin on the patient, as well as the angle at which they are taken. Without appropriate augmentation, deep learning models can erroneously overfit to data specific features instead of learning clinically relevant lesion features.

The augmented pipeline generated by the construction adds controlled geometric and photometric corrections which do not alter morphology of the lesions but consist of more robustness. Spatial transformations (horizontal flipping and shift-scale-rotate) enhance variation to both orientation and positional mismatches, but variations in brightness and contrast variations explain variations in illumination. The color changes caused by perturbation of hue and saturation are reminiscent of the changes in color that can be regularly seen during

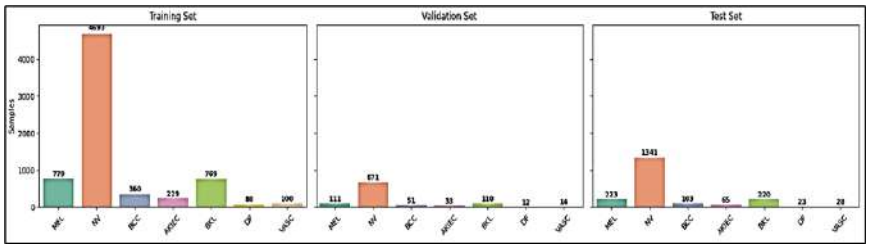

Figure 3. Training, Testing & Validation Split

```
FOR each model IN {Conv, EfficientNetV2, MobileNet} DO
    train_gen ← BalancedDataGen(train_data, batch_size, augmentation, model_preprocess)
    val_gen   ← ValDataGen(val_data, batch_size, model_preprocess)
    test_gen  ← TestDataGen(test_data, batch_size, model_preprocess)
END FOR
```

Figure 4. Pseudo-code of BalanceDataGen

dermoscopy practice, and contrast limited adaptive histogram equalization boosts contrast in the local area to better expose the fine structures of lesions.

The application of all augmentation operations is done stochastically when training and not when performing validation and testing to ensure that performance is evaluated without any bias. Given that the augmentation strategy can enhance the generalization ability of the training dataset to reassign previously unknown clinical situations, the augmentation strategy acts synergistically with the balanced sampling protocol by increasing both the functional diversity and the diagnostic semantics of the training dataset.

2.6 Deep Learning Architectures

The suggested architecture as shown in Figure 5 is a combination of three complementary convolutional neural network models, which are all pretrained on the ImageNet database.

2.6.1 ConvNeXt Tiny

ConvNeXt tiny is a current convolutional architecture that is inspired by the principles of transformer design. It uses corn kernel convolutions, depth wise separable layers and simplified normalization schemes thus supporting the robust top-down feature derivation process without distortion of the convolutional inductive biases.

2.6.2 EfficientNetV2-S

EfficientNetV2 -S is created to be more efficient in training and scale the parameters. The architecture implements fused MBConv blocks, scaling of depth and width optimization and better regularization, which makes it especially appropriate in medical image classification problems where data is scarce.

2.6.3 MobileNetV3-Large

MobileNetV3-Large is a small network that is capable of being deployed to resource-constrained devices. It makes use of inverted residual blocks, squeeze-and-excitation modules, and efficient nonlinearities, which allows its successful use in generating texture-sensitive feature extraction with a low amount of computational overhead.

2.7 Transfer Learning and Model Training

The two-stage transfer learning strategy is used to train each backbone network separately. The first step involves pretrained backbone layers that are frozen and the classification head is only trained to fit the high-level representations to dermoscopic images. Then all the layers are un-frozen to allow end-to-end fine-tuning. The Adam optimizer is used to implement training with an initial learning rate of 0.0001 ,dropout regularization and label smoothing to enhance the process of generalization and calibration. Principles used in the mitigation of training and overfitting with early stopping and learning-rate scheduling are implemented.

2.8 Ensemble Prediction and Explainability via Ensemble SmoothGrad

Final predictions are then made by probability-level ensemble averaging after the training of every single model. On every input image, ConvNeXt-Tiny, EfficientNetV2-S, and MobileNetV3-Large softmax probability outputs are averaged and the mean probability of the highest-ranked class is used as the final prediction. Such a fusion strat-

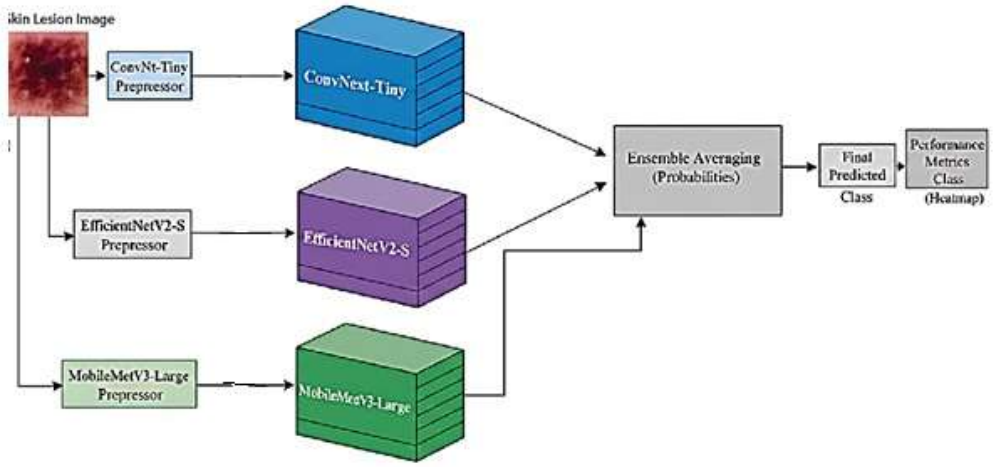

Figure 5. Proposed Deep Learning Architecture

egy minimizes the prediction variance, autocratic intentions in the individual models, and enhances strength, especially in small, not the majority classes of lesions. For a predicted class, SmoothGrad generates saliency by averaging gradients over noise-perturbed inputs:

$$S = \frac{1}{N}\sum_{n=1}^{N} \left|\nabla_x f_c\left(x + \epsilon_n\right)\right|. \tag{5}$$

The ensemble saliency map is computed as:

$$S_{\text{ens}} = \frac{1}{3}\left(S_{\text{convnext}} + S_{\text{effv2}} + S_{\text{mobilenet}}\right). \tag{6}$$

Saliency maps based on SmoothGrad are created on each of the trained models to guarantee interpretability by averaging gradients on inputs with noise perturbation. The saliency maps that are generated by the three models are then averaged to come up with an ensemble explanation. This method will emphasize predictable areas of lesion, including pigmentation patterns, irregular edges whereas reducing the background artifact making clinical confidence on the model predictions better.

3 Results and Discussion

The proposed three-stream ensemble was tested on HAM10000 data set based on the independent test set of 20 percent of the total image. The performance was measured based on the overall accuracy, class-wise precision, and recall, F1-score, and the confusion matrix. They report the results of individual models and the ensemble classifier to study the standalone and combined performance.

3.1 Performance Classification

The proposed hybrid ensemble model as summarized in Table 1 shows better classification performance than all the backbone networks. ConvNeXt-Tiny with the highest test accuracy of 89.37 percent and balanced macro-averaged precision, recall, and F1-score represents the model with the highest overall generalization. EfficientNetV2-S, and MobileNetV3-Large give competitive results using lower computational complexity, but their relatively low macro-level scores imply that they are insensitive to underrepresented classes of lesions. The hybrid ensemble based on probability-level averaging with the complementary feature representations of ConvNeXt-Tiny, EfficientNetV2-S, and MobileNetV3-Large achieves the maximum test accuracy (89.72), and significant improvements in the macro-averaged precision (0.8756) and F1-score (0.8551). As Table 1 shows, these findings confirm that the ensemble strategy is able to overcome individual model constraints, increase robustness and improve classification consistency across all lesion types.

3.2 Confusion Matrix Analysis

The confusion matrices in Fig.6, 7, 8 and 9 give a more in-depth understanding of how each of the individual models and the proposed ensemble will behave in terms of class prediction. The melanocytic nevi (NV) are always recognized with a high level of accuracy across all architectures, as they are overwhelming in the dataset and have specific visual features. Nonetheless,

Table 1. Performance Comparison of Models

Model	Acc. (%)	Prec.	Rec.	F1
ConvNeXt-Tiny	89.37	0.826	0.843	0.832
EfficientNetV2-S	86.82	0.808	0.781	0.793
MobileNetV3-Large	86.12	0.810	0.812	0.808
Proposed Model	89.72	0.876	0.840	0.855

the individual models are characterized by significant confusions in terms of visually similar classes, namely, melanoma (MEL) and benign keratosis-like lesions (BKL), basal cell carcinoma (BCC) and actinic keratoses (AKIEC). ConvNeXt-Tiny is somewhat balanced in its performance but has been misclassified MEL mostly as NV, whereas EfficientNetV2-S is more confused with AKIEC and BKL. MobileNetV3-Large is better at sensitizing to minority classes like dermatofibroma (DF) and vascular lesions (VASC) at the expense of confusion between MEL and NV. Conversely, the ensemble confusion matrix demonstrates a more vivid diagonal dominance and a better separation of the classes and less misclassification in most categories. It is important to note that the ensemble causes less confusion between clinically critical pairs like MELBKL and BCCAKEC, but it does not lower familiarity with few common classes such as DF and VASC. All of these observations confirm that probability-level ensemble fusion is an effective solution to the individual model weaknesses, to the issue of high-robustness to class imbalance, and to providing more clinically reliable classification results.

3.3 Training and Convergence Analysis

The validation accuracy-loss curves and the validation loss of ConvNeXt-Tiny, EfficientNetV2-S, and MobileNetV3-Large show that all models converge steadily and can generalize successfully. ConvNeXt-Tiny exhibits a high convergence rate, with the first few epochs showing rapid improvement in accuracy, and the validation accuracy reaching a high of 0.90. Training and validation errors decrease steadily, indicative of efficient learning with limited overfitting (see Figure 10 for ConvNeXt-Tiny). EfficientNetV2-S also shows fairly smooth convergence, achieving acceptable validation accuracy and reduction of validation loss with a steady increase (Figure 11 for EfficientNetV2-S). MobileNetV3-Large demonstrates slower convergence due to its lightweight structure, but the integration converges well, keeping the deviation between training and validation curves minimal, suggesting good generalization despite lower model capacity (Figure 12 for MobileNetV3-Large).

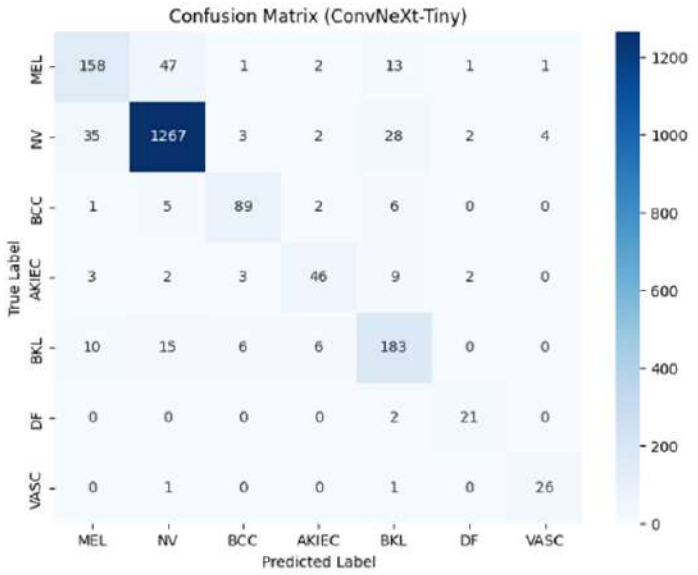

Figure 6. Confusion Matrix of ConvNeXt-Tiny

In all models, the moderate separation between training and validation measures shows that label smoothing, balanced sampling, and targeted augmentation effectively curb overfitting. The convergence trends validate that the proposed training strategy is well-tuned and supports robust feature learning for dermoscopic image classification.

3.4 SmoothGrad-Based Explainability Analysis

SmoothGrad-based visualizations of saliency provide a qualitative understanding of how the proposed models process images, indicating areas that contribute most to the predicted class. Each individual model produces a

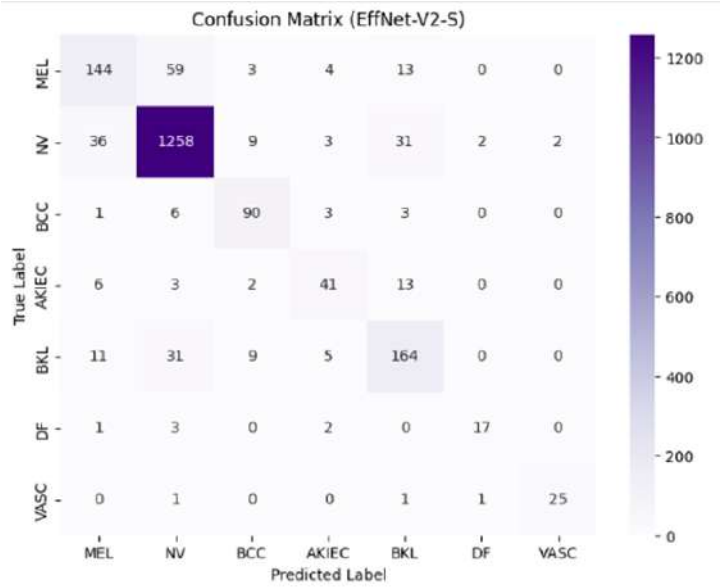

Figure 7. Confusion Matrix of EfficientNetV2-S

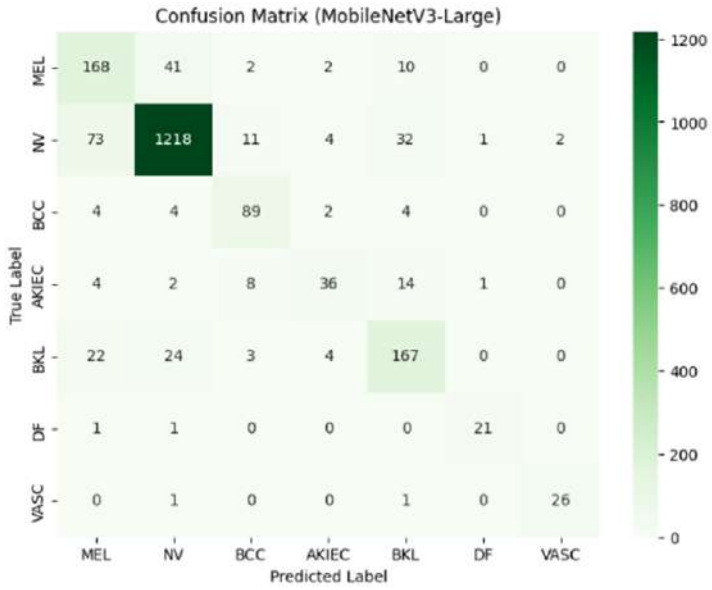

Figure 8. Confusion Matrix of MobileNetV3-Large

SmoothGrad map, shown in Figure **??**, by averaging gradients across multiple noise-perturbed versions of the input image. This approach produces sharper and more stable explanations than traditional gradient-based methods.

For ConvNeXt-Tiny, the saliency map highlights the lesion shape in the center and the edges of the image, indicating that the model focuses on hierarchical and global elements. EfficientNetV2-S captures a smaller attention region concentrated on high-contrast lesion areas, suggesting sensitivity to discriminative texture patterns. MobileNetV3-Large emphasizes smaller-scale texture and color differences over a larger lesion size, consistent with its lightweight local feature extraction structure. The SmoothGrad visualization resulting from averaging all three models shows the most consistent and clinically significant attention pattern, focusing on the lesion core, peripheral margins, and vascular regions while avoiding background artifacts. This cross-architectural cooperation demonstrates that the ensemble is not only more effective in classification but also provides higher-quality explanations that align with dermatological diagnostic criteria, increasing clinical confidence and transparency.

According to the recent findings as Table 2. shows on the automated classification of skin lesions based on the HAM10000 dataset, our proposed ensemble model outperforms all the models obtaining an accuracy of 89.7%. Previous studies conducted by Mahbod et al. (2023) used a multi-Cnn ensemble model, which gave an accuracy of 86.2%. After that, Thwin S.M. et al. (2024) trained heterogeneous deep systems (VGG16, Inception-V3, and ResNet50) and applied them to a deep ensemble system, which increased the performance to 87.9%. Recently, a fuzzy rank-based deep ensemble based on a combination of Xception, InceptionResNetV2, and MobileNetV2 and having the highest recorded accuracy of 89.3, was presented by Halder A. et al. (2025). These findings suggest that a wide variety of model fusion and smart decision aggregation have a considerable positive impact on the classification in the analysis of dermoscopic images.

3.5 Qualitative Prediction Analysis

Figure 14 shows a representative test image that is rightfully diagnosed with Basal Cell Carcinoma (BCC). A posterior probability of the BCC class in the model (0.825) is significantly higher than the probabilities of the clinically similar classes, including Melanoma (0.0297) and Benign Nevus (0.0628), thus demonstrating a robust discriminative performance of the model.

4 Conclusion and Future Work

This paper proposed a strong and explainable deep learning model of automated multiclass skin lesion

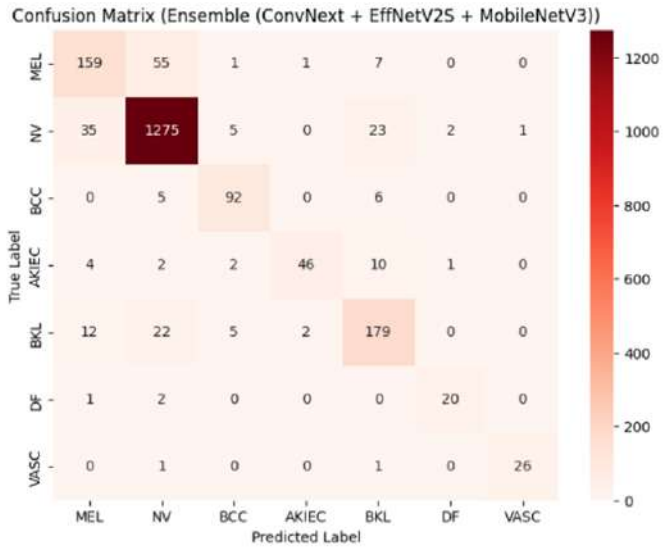

Figure 9. Confusion Matrix of Proposed Ensemble Model

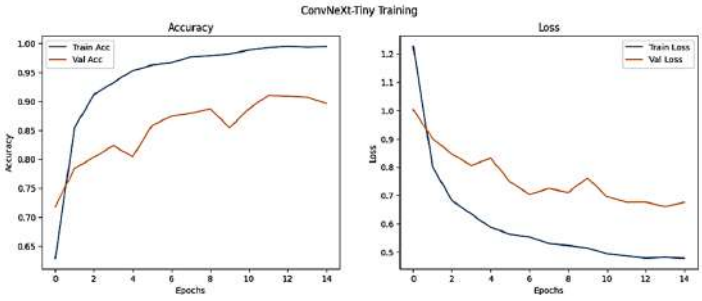

Figure 10. Training accuracy and loss of ConvNext-Tiny Model

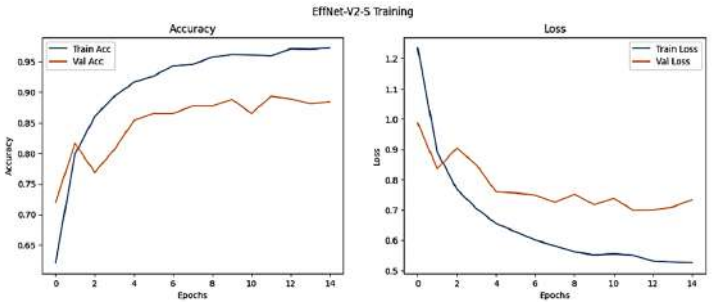

Figure 11. Training accuracy and loss of EfficientnetV2-S Model

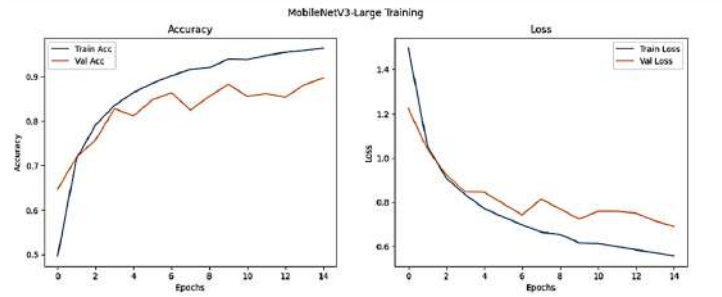

Figure 12. Training accuracy and loss of MobileNetV3-Large Model

classification based on dermoscopic pictures of the HAM10000 dataset. The proposed approach, through the combined use of three complementary convolutional architectures, such as ConvNeXt-Tiny, EfficientNetV2-S, and MobileNetV3-Large, successfully reduced the major issues in the analysis of dermatological images, such as extreme bias in classes, high similarity between them, scarcity of data, and the necessity to interpret the results clinically.

The sampling strategy was designed to provide the same representation of the classes in the training without depending on the synthetic oversampling or loss reweighting mechanism, and maintain the natural distribution of the lesion features. Data augmentation on the Albumentations was done selectively to further generalization and decrease overfitting. Ensemble-based fusion at the probability level tapped into the complementary ability of the separate models leading to more robustness, as well as improved macro-averaged performance, especially of underrepresented lesion classes. The experimental findings proved that the hybrid ensemble obtained the best overall test accuracy and macro F1-score than individual backbone models.

Besides quantitative performance improvements, stable and clinically significant explanations of model predictions were also obtained through incorporation of SmoothGrad-based ensemble saliency visualization. The collective explanations were always restricted to the areas of the lesion that were diagnostically relevant (like: pigmentation patterns, pigmentation borders, and vascular structures), whereas other background artifacts were suppressed. This interpretability is vital in terms of clinical acceptance and helps to deploy the suggested framework as a decision-support tool in practice in dermatology.

In general, the suggested system provides a desirable accuracy, robustness, interpretability, and computational efficiency ratio, which can be deployed to low-resource clinical settings and may be used in

Figure 13. SmoothGradCam showing explainabilty of models on the affected area

Table 2. Accuracies Comparison with Previous Studies

Ref.	Year	Model	Acc. (%)
[22]	2024	Multi-CNN Ensemble	86.2
[23]	2024	VGG16 + Inc.-V3 + ResNet50 Ens.	87.9
[24]	2025	Xception + IncRes-NetV2 + Mo-bileNetV2 Fuzzy Ens.	88.9
Proposed	2025	ConvNeXt-T + EffNetV2-S + Mo-bileNetV3 Ens.	89.7

Figure 14. Qualitative Prediction Analysis

possible mobile or edge-AI screening services. Even though the findings may be promising, there are still a number of future research directions. First, it might be further improved by adding explicit lesion segmentation or attention-based mechanisms that would better distinguish between similarly looking lesions, e.g. melanoma and benign keratosis-like lesions. Second, the framework could be expanded to incorporate clinical metadata (e.g. the age of a patient, the placement of his or her lesions, or the medical history) and theoretically a greater number of factors could be used to boost the accuracy of diagnosis and its relevance.

There is also the possibility that future research can develop hybrid CNN-transformer networks to have enhanced power to capture long-range contextual dependencies without compromising the computational efficiency. Furthermore, it would be beneficial to test the suggested framework on bigger and more heterogeneous multi-center datasets to determine how it can be applied to other populations and imaging cases in general. Deployment wise, additional optimization of real-time inference and validation in prospective clinical studies would be the next steps leading to real-world implementation. Lastly, further development of explainability and uncertainty estimation and clinician-in-the-loop evaluation may contribute to the improvement of trust and usability, thus facilitating reliable and transparent AI-assisted dermatologic diagnosis.

Author Contributions

Muhammad Faris: Conceptualization, Methodology, and supervision. **Falak Khalid**: Data curation, Software, Writing- Original draft preparation. **Maida Shahid**: Visualization and investigation. **Zeeshan Ul Haque**: Software Validation, Writing-Reviewing and Editing. **Syed**

Ibad Hasnain: Supervision **Mir Farooq Ali**: Validation of the concept, and Writing- Reviewing.

Compliance with Ethical Standards

It is declare that all authors don't have any conflict of interest. It is also declare that this article does not contain any studies with human participants or animals performed by any of the authors.

References

[1] A. H. Roky *et al.*, "Overview of skin cancer types and prevalence rates across continents," *Cancer Pathogenesis and Therapy*, vol. 3, no. 2, pp. 89–100, 2025.

[2] B. Ahmed, M. I. Qadir, and S. Ghafoor, "Malignant melanoma: skin cancer—diagnosis, prevention, and treatment," *Critical Reviews in Eukaryotic Gene Expression*, vol. 30, no. 4, 2020.

[3] S. Levit, J. Shoykhet, and E. Levit, "Comprehensive insights into basal cell carcinoma: causes, presentation, prevention, and modern therapeutic approaches," *Cancer Medicine*, vol. 14, no. 24, p. e71448, 2025.

[4] G. P. Pfeifer, "Mechanisms of UV-induced mutations and skin cancer," *Genome Instability & Disease*, vol. 1, no. 3, pp. 99–113, 2020.

[5] M. Abubakar, "Overview of skin cancer and risk factors," *International Journal of General Practice Nursing*, vol. 2, no. 3, pp. 42–56, 2024.

[6] L. Rey-Barroso, S. Peña-Gutiérrez, C. Yáñez, F. J. Burgos-Fernández, M. Vilaseca, and S. Royo, "Optical technologies for the improvement of skin cancer diagnosis: a review," *Sensors*, vol. 21, no. 1, p. 252, 2021.

[7] N. Melarkode, K. Srinivasan, S. M. Qaisar, and P. Plawiak, "AI-powered diagnosis of skin cancer: a contemporary review, open challenges and future research directions," *Cancers*, vol. 15, no. 4, p. 1183, 2023.

[8] N. Razmjooy *et al.*, "Computer-aided diagnosis of skin cancer: a review," *Current Medical Imaging Reviews*, vol. 16, no. 7, pp. 781–793, 2020.

[9] S. Haque, F. Ahmad, V. Singh, D. M. Mathkor, and A. Babegi, "Skin cancer detection using deep learning approaches," *Cancer Biotherapy & Radiopharmaceuticals*, vol. 40, no. 5, pp. 301–312, 2025.

[10] Muhammad Faris, Ramsha Qayyum, Crescenzo Pepe, Mir Farooq Ali, and Silvia Maria Zanoli, "Multi-Class Skin Disease Detection Using Deep Learning Hybrid Method," in *2025 26th International Carpathian Control Conference (ICCC)*, IEEE, 2025, pp. 1–5.

[11] S. Adamu *et al.*, "The future of skin cancer diagnosis: a comprehensive systematic literature review of machine learning and deep learning models," *Cogent Engineering*, vol. 11, no. 1, p. 2395425, 2024.

[12] S. J. Kumar, G. P. Kanna, D. P. Raja, and Y. Kumar, "A comprehensive study on deep learning models for the detection of ovarian cancer and glomerular kidney disease using histopathological images," *Archives of Computational Methods in Engineering*, vol. 32, no. 1, pp. 35–61, 2025.

[13] C. N. Kanani and P. U. Jadeja, "Skin disease diagnosis: a comprehensive survey of models, datasets, and clinical applications," *Cureus*, vol. 2, no. 1, 2025.

[14] B. Kamaraj *et al.*, "Explainable AI for skin cancer detection using convolutional neural networks with SE and Grad-CAM," *Journal of Investigative Dermatology*, vol. 145, no. 8, p. S21, 2025.

[15] H. L. Gururaj, N. Manju, A. Nagarjun, V. M. Aradhya, and F. Flammini, "DeepSkin: a deep learning approach for skin cancer classification," *IEEE Access*, vol. 11, pp. 50205–50214, 2023.

[16] M. Benaly, B. Kouach, I. Alihamidi, and L. Hlou, "Explainable AI for skin cancer classification: unlocking insights with Grad-CAM and Grad-CAM++," in *Proc. Int. Conf. on Circuit, Systems and Communication (ICCSC)*, IEEE, 2025, pp. 1–4.

[17] T. Öznacar and N. Varol Kayapunar, "Advanced skin cancer prediction with medical image data using MobileNetV2 deep learning and optimized techniques," *Scientific Reports*, vol. 15, no. 1, p. 28962, 2025.

[18] E. Zhao, "Deep learning models-based skin cancer classification and early detection," in *ITM Web of Conferences*, vol. 78, p. 02007, 2025.

[19] A. V. Dahat, U. T. Kute, T. R. Mahore, A. A. Joshi, S. Dhanraj, and U. D. Pande, "Detection of multi-class skin cancer using stochastic gradient descent augmentation model and activation mapping," *Journal of Innovative Image Processing*, vol. 7, no. 4, pp. 1415–1435, 2025.

[20] J. V. Tembhurne, N. Hebbar, H. Y. Patil, and T. Diwan, "Skin cancer detection using ensemble of machine learning and

deep learning techniques," *Multimedia Tools and Applications*, vol. 82, no. 18, pp. 27501–27524, 2023.

[21] P. Tschandl, "The HAM10000 dataset: a large collection of multi-source dermatoscopic images of common pigmented skin lesions," Harvard Dataverse, 2018, doi: 10.7910/DVN/DBW86T.

[22] A. H. Efat, S. M. Hasan, M. P. Uddin, and M. A. Mamun, "A multi-level ensemble approach for skin lesion classification using customized transfer learning with triple attention," *PLOS ONE*, vol. 19, no. 10, p. e0309430, 2024.

[23] S. M. Thwin and H.-S. Park, "Skin lesion classification using a deep ensemble model," *Applied Sciences*, vol. 14, no. 13, p. 5599, 2024.

[24] A. Halder, A. Dalal, S. Gharami, M. Woźniak, M. F. Ijaz, and P. K. Singh, "A fuzzy rank-based deep ensemble methodology for multi-class skin cancer classification," *Scientific Reports*, vol. 15, no. 1, p. 6268, 2025, doi: 10.1038/s41598-025-90423-3.