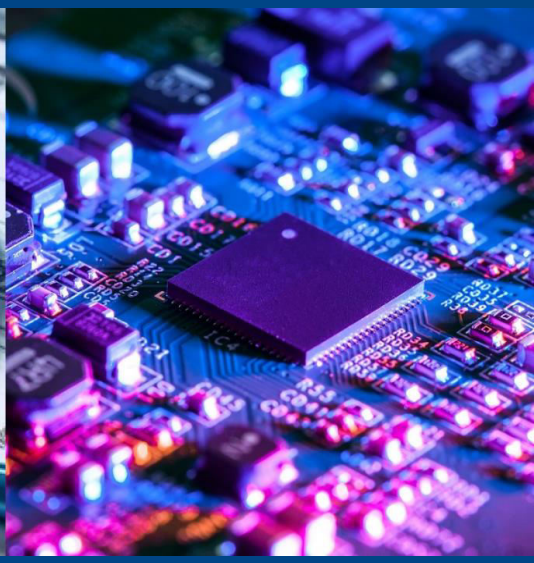
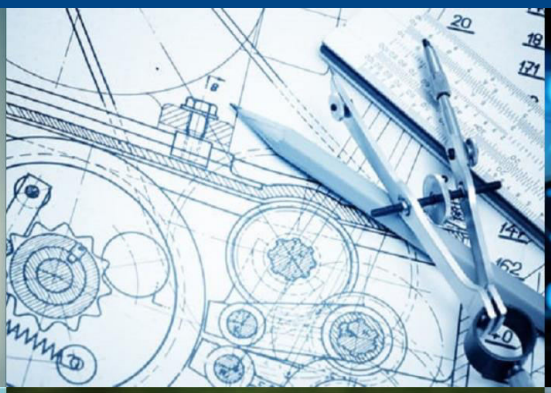
ISSN: 2582-7219

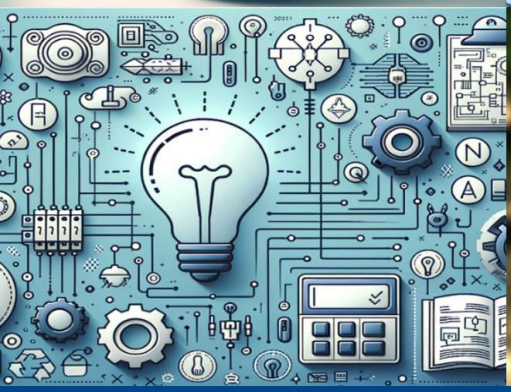




International Journal of Multidisciplinary Research in Science, Engineering and Technology

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 8.206 **Volume 9, Issue 4, April 2026**

© 2026 IJMRSET | Volume 9, Issue 4, April 2026| DOI:10.15680/IJMRSET.2026.0904460

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Recall-Oriented Deep Learning Framework for Skin Cancer Detection with Explainability and Fairness Analysis

Aniruddha S Rumale[1], Dipak D Bage[2], Devvrat S Kokane[3], Kaivalya G Aher[4], Pankaj N Rathod[5], Sarang K Phad[6]

Professor, Dept. of Information Technology, Sandip Institute of Technology and Research Centre, Nashik, India[1]

Associate Professor & HOD, Dept. of Information Technology, Sandip Institute of Technology and Research Centre, Nashik, India[2]

UG Scholars, Dept. of Information Technology, Sandip Institute of Technology and Research Centre, Nashik, India[3,4,5,6]

ABSTRACT: Skin lesion classification from dermoscopic images remains a challenging problem due to the high visual similarity between benign and malignant lesions, significant class imbalance, and variations in imaging conditions. Early and accurate detection of skin cancer is essential for effective treatment; however, manual diagnosis is often difficult, especially in the early stages where visual cues are subtle. In this work, a deep learning-based framework is proposed for automated skin lesion classification using a transfer learning approach. The system incorporates a carefully designed training and inference pipeline aimed at improving generalization and minimizing clinically critical errors, particularly false negatives. To enhance sensitivity toward malignant cases, a recall-oriented threshold optimization strategy is applied. Model interpretability is addressed through Grad-CAM++, enabling visualization of discriminative regions influencing predictions. In addition, fairness across different skin tones is analyzed using an ITA-based subgroup evaluation, providing insights into potential performance disparities. Experimental evaluation on the ISIC dataset demonstrates competitive performance, achieving a ROC-AUC of 0.926 and a recall of 0.879. The results indicate that the proposed framework offers a reliable, interpretable, and clinically relevant solution for automated skin lesion classification, with improved sensitivity and consideration for fairness.

KEYWORDS: Skin cancer detection, dermoscopic image analysis, deep learning, transfer learning, ResNet50, explainable AI, Grad-CAM++, fairness analysis, threshold optimization, medical image classification.

I. INTRODUCTION

Skin is the largest organ of the human body and plays a vital role in protecting against environmental factors such as ultraviolet (UV) radiation, pathogens, and physical injury. However, prolonged exposure to harmful elements, particularly UV radiation, can lead to abnormal cell growth and the development of skin cancer. Among the various types, melanoma is the most aggressive and life-threatening form. Although it accounts for a smaller proportion of total cases, it is responsible for the majority of skin cancer-related deaths, making early detection critically important [1].

Diagnosing skin lesions accurately remains a challenging task, even for experienced dermatologists. Benign and malignant lesions often share similar visual characteristics, with subtle differences in color, texture, and structure. This makes manual assessment difficult and prone to variability. Traditional diagnostic methods, including visual inspection and dermoscopy, rely heavily on clinical expertise. Studies indicate that dermatologists achieve diagnostic accuracies of around 75–80%, highlighting both the complexity of the task and the possibility of misclassification [2], [3]. While the ABCD rule—based on asymmetry, border irregularity, color variation, and diameter—provides a useful guideline, it is not always sufficient for reliable diagnosis in complex or ambiguous cases [4].

To overcome these limitations, computer-aided diagnostic (CAD) systems have been developed to assist clinicians in analyzing dermoscopic images. Recent advances in machine learning, particularly deep learning, have significantly improved the performance of such systems. Convolutional neural networks (CNNs) have emerged as a dominant

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

approach due to their ability to automatically learn hierarchical features directly from image data [5]. Additionally, transfer learning, which leverages models pre-trained on large-scale datasets such as ImageNet, has enabled effective model training even with limited medical data [6].

Despite these advancements, several important challenges remain. Most existing studies emphasize overall accuracy or ROC-AUC, often overlooking clinically critical errors such as false negatives. In medical diagnosis, the impact of errors is not uniform: missing a malignant case can delay treatment and pose serious risks to patient health, whereas false positives generally lead to additional tests but are less severe. Therefore, designing systems that prioritize the detection of malignant lesions and minimize false negatives is essential.

Another key challenge is the lack of interpretability in deep learning models. Although CNN-based approaches can achieve high performance, they often behave as black-box systems, making it difficult to understand the reasoning behind their predictions. Techniques such as Grad-CAM have been introduced to provide visual explanations by highlighting important regions in the input image [7]. However, these methods can sometimes produce diffuse or misleading attention maps, especially when models rely on irrelevant features such as hair artifacts or background patterns.

In addition, generalization remains a major concern. Models trained on a specific dataset often perform poorly when applied to new datasets due to domain shift. Variations in imaging devices, acquisition conditions, and patient demographics can significantly affect performance [8]. This underscores the need for robust evaluation strategies and adaptive mechanisms that improve model reliability across diverse data distributions.

In this work, we present a recall-oriented deep learning framework for binary skin lesion classification that addresses these challenges. The proposed approach is based on a ResNet50 architecture with transfer learning and incorporates strategies to enhance generalization while reducing false negatives. A recall-focused threshold optimization technique is introduced to improve sensitivity to malignant cases. Furthermore, Grad-CAM++ is utilized to provide more precise visual explanations, along with a refinement step to enhance localization quality. The model is evaluated on both internal and external datasets to assess robustness, and threshold recalibration is applied to maintain performance under domain shift.

The main contributions of this study are summarized as follows:

1. A classification framework designed to prioritize recall and reduce false negatives in skin cancer detection.
2. A threshold optimization strategy that adapts decision boundaries based on data characteristics.
3. An explainability approach using Grad-CAM++ with refinement for improved localization.
4. A comprehensive evaluation across multiple datasets to assess robustness and generalization.

Overall, the proposed framework aims to bridge the gap between high-performing deep learning models and their practical clinical applicability, providing a reliable, interpretable, and sensitivity-focused solution for early skin cancer detection.

II. LITERATURE REVIEW

In recent years, considerable research has focused on the automated detection and classification of skin cancer using machine learning and deep learning techniques. The growing incidence of skin cancer worldwide has increased the demand for accurate and reliable computer-aided diagnostic (CAD) systems to support dermatologists in early diagnosis and clinical decision-making [9]. Earlier approaches primarily relied on handcrafted feature extraction combined with traditional classifiers; however, these methods were limited in their ability to capture complex visual patterns present in dermoscopic images.

With the advancement of deep learning, convolutional neural networks (CNNs) have become the dominant approach for skin lesion classification. Architectures such as VGGNet [10], ResNet [11], DenseNet [12], MobileNet [13], and EfficientNet [14] have demonstrated strong capability in learning hierarchical feature representations and improving classification performance. Several studies have proposed enhancements to these architectures, including hybrid CNN frameworks and multi-model systems that combine different feature extractors to better capture lesion characteristics [15], [16]. In addition, multi-stage pipelines integrating CNNs with classifiers such as Support Vector Machines (SVMs) have been explored to further improve prediction accuracy [17].

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Data augmentation plays a crucial role in addressing the limited size and imbalance of medical image datasets. Common techniques include geometric transformations such as rotation and flipping, as well as color adjustments to simulate variations in imaging conditions [18]. More advanced approaches involve synthetic data generation using Generative Adversarial Networks (GANs) and oversampling techniques such as SMOTE and ADASYN [19]. While these methods can improve generalization and reduce class imbalance, their effectiveness is often dependent on dataset characteristics and does not always lead to consistent performance gains [20].

The ISIC (International Skin Imaging Collaboration) dataset has become a widely used benchmark for evaluating skin lesion classification models. Many studies have applied transfer learning with pre-trained models such as InceptionV3, VGG16, DenseNet, and AlexNet, achieving varying levels of performance depending on architecture and training strategy [21], [22]. However, despite high reported accuracy, models often struggle to generalize across datasets due to domain shift.

Class imbalance remains a significant challenge in this domain. In most datasets, benign samples substantially outnumber malignant cases, leading to biased models that favor the majority class. To mitigate this issue, researchers have explored cost-sensitive learning, weighted loss functions, and other optimization strategies to improve sensitivity toward malignant lesions [23].

Ensemble learning has also been widely adopted to enhance performance. By combining multiple models, ensemble approaches reduce prediction variance and leverage complementary strengths of individual architectures. Techniques such as stacking, weighted averaging, and multi-scale ensembles have shown improved robustness compared to single-model approaches [24], [25].

Despite these advances, several limitations persist. Many existing works focus primarily on improving overall accuracy, often neglecting clinically critical aspects such as minimizing false negatives. Furthermore, interpretability remains limited, as deep learning models often operate as black-box systems. In addition, generalization across datasets continues to be a challenge due to variations in imaging conditions and data distributions. A brief comparison of representative methods highlights the limitations of existing approaches.

Compared to prior approaches, the proposed framework emphasizes recall-oriented performance to reduce missed malignant cases, integrates explainability through Grad-CAM++, and evaluates robustness across datasets. These design choices address key limitations in existing methods and improve the practical applicability of automated skin lesion classification systems.

III. METHODOLOGY OF PROPOSED SURVEY

This study presents a binary skin lesion classification framework based on transfer learning using a deep convolutional neural network [6]. The primary objective is to accurately distinguish between benign and malignant lesions while minimizing false negatives and improving model interpretability. The proposed pipeline consists of data preprocessing, model training using a ResNet50 backbone, threshold optimization, and evaluation under both internal and external settings. Preprocessing includes resizing, normalization, and data augmentation to enhance generalization, particularly for imbalanced dermoscopic datasets. The model leverages ImageNet pre-trained weights with a modified classification head tailored for binary prediction. To address clinically critical errors, a recall-oriented threshold optimization strategy is applied. Interpretability is incorporated using Grad-CAM++, enabling visualization of important regions influencing predictions. The framework is further evaluated on an external dataset to assess robustness under domain shift, and threshold recalibration is applied to maintain performance consistency. Additionally, fairness-aware evaluation is conducted to analyze performance across different skin tone groups [23]. Overall, the proposed framework integrates model optimization, threshold calibration, explainability, and fairness analysis into a unified system for reliable skin cancer detection.

© 2026 IJMRSET | Volume 9, Issue 4, April 2026| DOI:10.15680/IJMRSET.2026.0904460

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

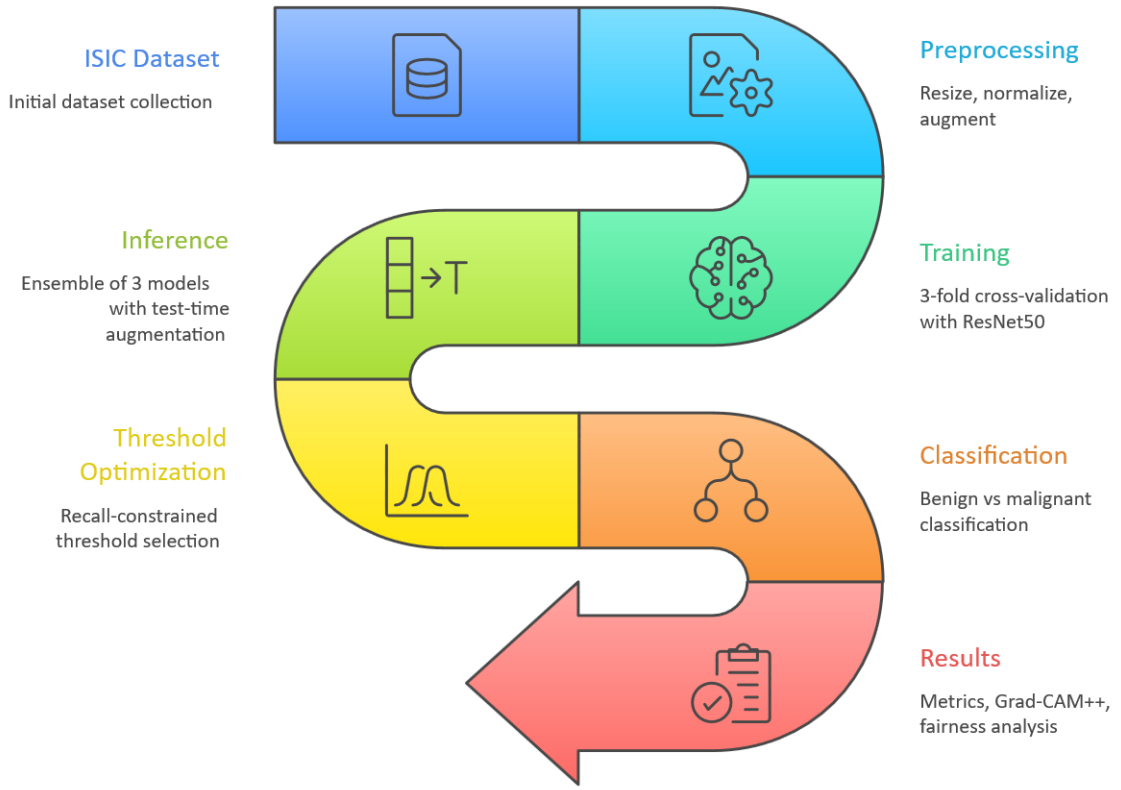


Fig. 1. Overview of the proposed skin lesion classification pipeline from data preprocessing to model training, inference, and evaluation

Data Analysis

All experiments are conducted on the ISIC 2018 dermoscopy dataset, a widely used benchmark for automated skin lesion analysis [21]. The dataset consists of high-resolution RGB images collected from multiple clinical sources, capturing variations in imaging conditions, acquisition devices, and lesion characteristics [3]. This diversity makes it suitable for evaluating the robustness of deep learning models. For this study, a binary classification setup is adopted by grouping lesion categories into benign and malignant classes, reflecting a practical screening scenario. Representative samples from both classes are shown in **Fig. 2**. Benign lesions generally exhibit more regular shapes and uniform color patterns, whereas malignant lesions tend to display irregular borders, asymmetry, and heterogeneous pigmentation. However, visual overlap between the two classes makes the classification task challenging.

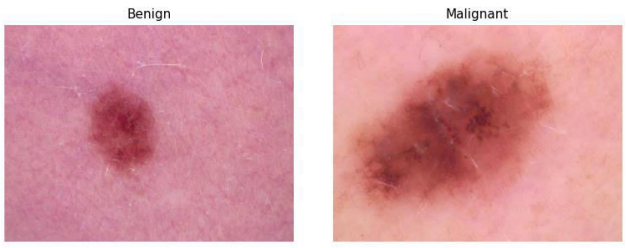


Fig. 2. Sample images of benign and malignant skin lesions.

© 2026 IJMRSET | Volume 9, Issue 4, April 2026| DOI:10.15680/IJMRSET.2026.0904460

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

The dataset is provided with predefined training, validation, and test splits, as summarized in **Table 2.** The class distribution is further illustrated in **Fig. 3**, highlighting a significant imbalance between benign and malignant samples. Malignant lesions constitute approximately 18.5% of the training data, corresponding to a ratio of roughly 4.4:1.

Table 1: Dataset Partition Statistics

Split	Total Images	Benign	Malignant	Benign Ratio	Malignant Ratio
Train	9,885	8,070	1,815	81.6%	18.4%
Validation	193	140	53	72.5%	27.5%
Test	1,493	1,205	288	80.7%	19.3%

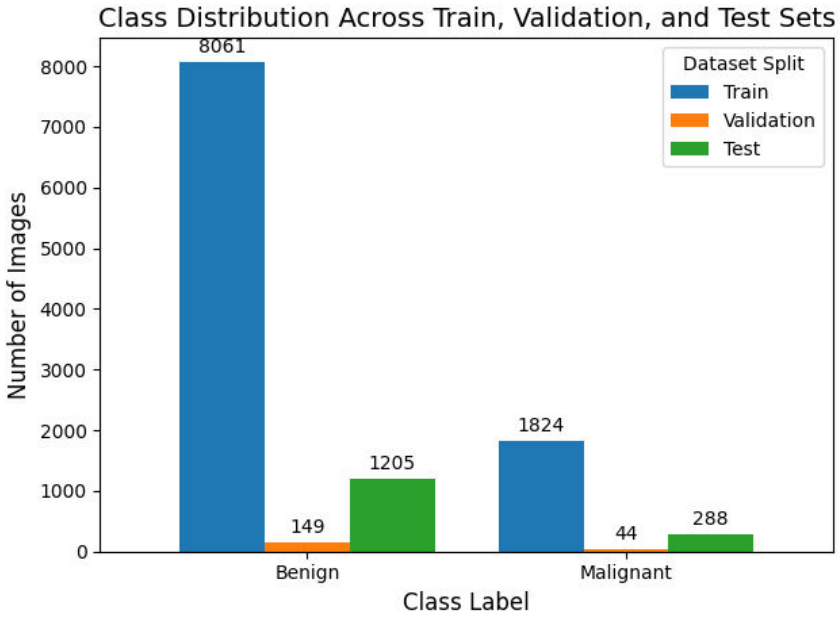


Fig. 3. Distribution of benign and malignant samples across training, validation, and test sets.

This imbalance reflects real-world prevalence but introduces challenges during training, as models may become biased toward the majority class. Therefore, specific strategies are required to ensure adequate sensitivity to malignant cases. Overall, the dataset presents a realistic and challenging scenario due to class imbalance, visual similarity between categories, and variability in acquisition conditions.

Preprocessing and Data Augmentation

All images were resized to 299×299 pixels before being fed into the model. This resolution provides a balance between preserving spatial details and maintaining computational efficiency, while remaining compatible with the selected backbone architectures. Pixel values were normalized using ImageNet channel-wise statistics (mean: 0.485, 0.456, 0.406; standard deviation: 0.229, 0.224, 0.225) to align with pre-trained feature representations. To improve generalization and reduce overfitting, data augmentation was applied exclusively during training [18]. Validation and test images were processed using only resizing and normalization to ensure unbiased evaluation. The augmentation strategy incorporates clinically relevant transformations. Random rotations and horizontal/vertical flips account for the absence of a fixed orientation in dermoscopic images. Color jitter simulates variations in imaging conditions across devices and clinical environments [18]. Random erasing improves robustness to occlusions caused by artifacts such as hair and markers, which may otherwise act as confounding factors [19].

© 2026 IJMRSET | Volume 9, Issue 4, April 2026| DOI:10.15680/IJMRSET.2026.0904460

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Backbone Selection

A controlled comparison was performed between ResNet50, EfficientNet-B3, and DenseNet, which are commonly used architectures in dermoscopic image analysis [11], [12], [14]. The goal was to identify the backbone that offers the best balance between generalization and clinically relevant performance.

All models were trained under identical conditions, including input resolution, augmentation strategy, optimization settings, and threshold selection procedure. This ensures that differences in performance are attributable solely to the backbone architecture. The results, summarized in **Table .2**, show that EfficientNet-B3 achieves the highest validation AUC but exhibits a larger gap between validation and test performance, indicating weaker generalization. DenseNet demonstrates the smallest gap, suggesting stable generalization; however, it underperforms in terms of recall and missed malignant cases. In contrast, ResNet50 achieves a balanced performance with high recall and a relatively small generalization gap. Based on these observations, ResNet50 is selected as the backbone for the proposed framework.

Table 2: Backbone comparison under identical training conditions

Model	Test AUC	Val AUC	Val→Test AUC gap	Test Recall	FN (missed malignant)	FP (false alarm)	Early stop epoch
EfficientNet-B3	**0.9144**	**0.9568**	**0.0424**	**0.7847**	**62**	**164**	**15**
DenseNet	**0.8980**	**0.9081**	**0.0101**	**0.6909**	**89**	**147**	**5**
ResNet50	**0.9159**	**0.9420**	**0.0261**	**0.8194**	**52**	**227**	**11**

Model Architecture

The final model is based on ResNet50 initialized with ImageNet pre-trained weights [6]. The architecture consists of convolutional layers organized into residual blocks, enabling efficient gradient propagation and stable training [11]. Input images of size 299×299×3 are processed to extract hierarchical feature representations. A global average pooling layer reduces spatial dimensions, producing a compact feature vector. The original ImageNet classification head is replaced with a task-specific head consisting of a dropout layer (p = 0.4) followed by a fully connected layer mapping to two output classes (benign and malignant). This modification enables effective adaptation to the binary classification task while reducing overfitting. The overall architecture of the proposed model is illustrated in **Fig. 4.**

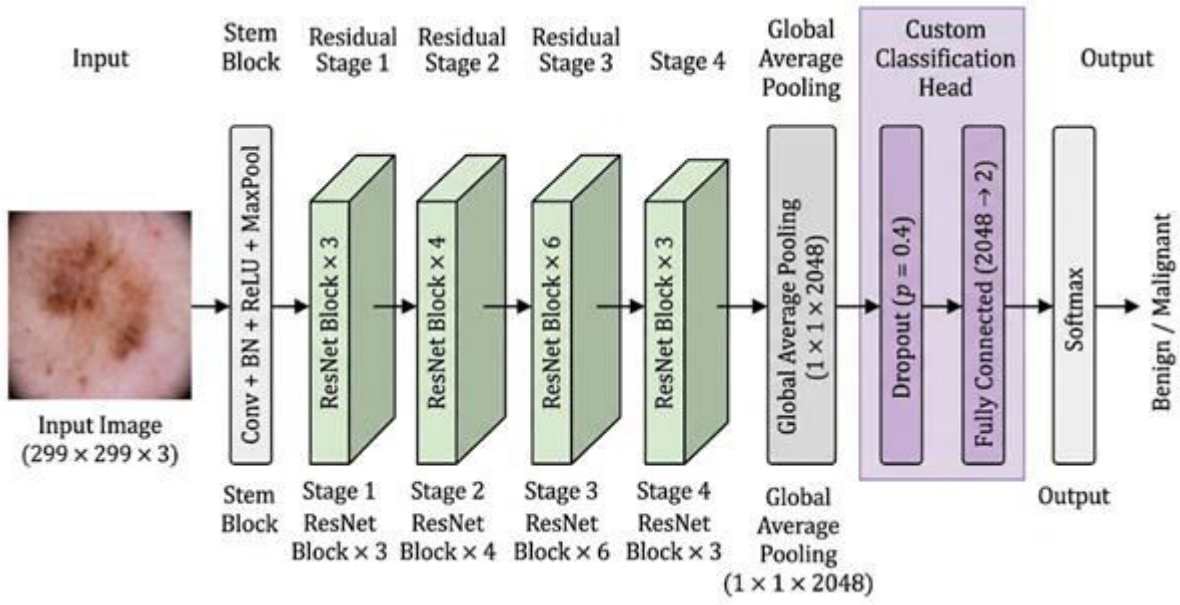


Fig. 4. Network architecture of ResNet50.

The architecture can be expressed as:

$$\mathbf{f(x) = W \cdot Drop\ (GAP(\Phi(x))) + b}$$

where Φ(x) represents feature extraction, GAP denotes global average pooling, and W and b are learnable parameters.

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Training Strategy

The model is trained using transfer learning with differential learning rates applied across the network. Lower learning rates are used for earlier layers to preserve general features, while higher rates are assigned to deeper layers and the classification head for task-specific learning. To address class imbalance, weighted cross-entropy loss is used, with class weights computed based on data distribution [23]. Label smoothing is applied to improve probability calibration and reduce overconfidence. Optimization is performed using the Adam optimizer with weight decay. A cosine annealing scheduler is used to gradually reduce the learning rate, ensuring stable convergence [20]. Early stopping based on validation loss (patience = 7 epochs) is applied, and the best-performing model is retained.

Evaluation and Inference Strategy

To ensure reliable evaluation, 3-fold stratified cross-validation is applied, maintaining class distribution across all folds [24]. This results in three independently trained models. During inference, predictions are combined using ensemble averaging, which reduces variance and improves stability [25]. Test-time augmentation (TTA) is also applied by evaluating multiple augmented versions of each image, and averaging their predictions [18]. This combined approach enhances robustness and provides more reliable performance compared to single-model inference.

Threshold Optimization

A fixed decision threshold (0.5) is often suboptimal in clinical settings where false negatives carry higher risk. Therefore, a recall-oriented threshold optimization strategy is employed. Thresholds are evaluated on the validation set in increments of 0.02. For each threshold, recall, precision, and F1-score are computed. The optimal threshold is selected by maximizing F1-score under a recall constraint [23]:

$$t^* = \arg\max_t \{F_1(t) \mid \text{Recall}(t) \geq \gamma\}$$

This ensures high sensitivity while maintaining a balanced precision–recall trade-off. The optimal threshold is found to be 0.46 for the default setting and 0.38 for fairness calibration.

Grad-CAM++ Explainability

Grad-CAM++ is used to provide visual explanations of model predictions [7]. It generates class-specific heatmaps by weighting feature maps using gradient information, enabling more precise localization than standard Grad-CAM. . The overall Grad-CAM++ pipeline is illustrated in **Fig. 5.**

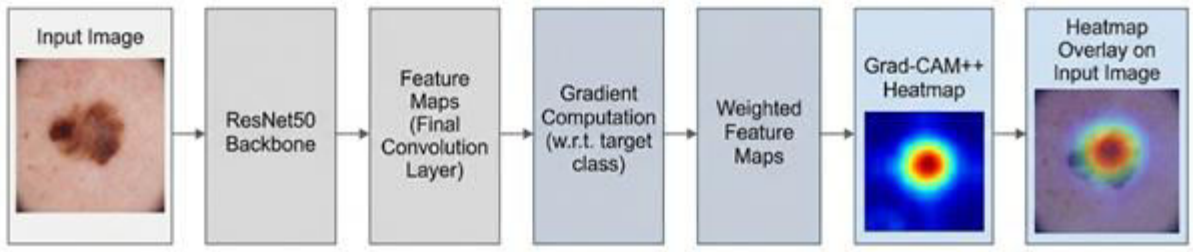


Fig. 5. Overview of the Grad-CAM++ pipeline for generating class-discriminative heatmaps from convolutional feature maps.

The heatmap is computed as:

$$L^c_{\text{Grad-CAM++}} = \text{ReLU}(\sum_k \alpha_k^c \cdot A^k)$$

where A^k represents activation maps and α_k^c denotes corresponding weights.

The final convolutional layer (layer4[-1]) is used as the target layer. A sharpening step retains the top 25% of activations to focus on the most relevant regions. The resulting heatmaps are overlaid on input images, highlighting regions associated with malignant features.

Fairness Analysis

Fairness is evaluated using the Individual Typology Angle (ITA) as a proxy for skin tone classification [28]. ITA is computed from LAB color space values extracted from non-lesion regions to avoid bias. Images are grouped into light (ITA > 28) and dark (ITA ≤ 28) categories. Model performance is evaluated separately for each group to identify

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

potential disparities. Additionally, threshold calibration is applied to assess whether performance gaps can be reduced without retraining the model.

IV. RESULTS AND DISCUSSION

Implementation Details

All experiments were conducted in Python using the PyTorch framework. Training was performed on an NVIDIA Tesla T4 GPU with a batch size of 32. The model was optimized using the Adam optimizer with weight decay. Differential learning rates were applied across network layers, and a cosine annealing scheduler was used to ensure stable convergence. Early stopping based on validation loss (patience = 7 epochs) was employed to prevent overfitting. A fixed random seed was used to ensure reproducibility. The 3-fold cross-validation pipeline required approximately 45 minutes per fold, resulting in a total training time of around 150 minutes. Inference on the test set, including ensemble prediction with test-time augmentation, was completed in approximately 12 minutes.

Classification Performance

The final ensemble model, consisting of three cross-validation models combined with test-time augmentation, was evaluated on the held-out test set at the default threshold (t = 0.46). The performance metrics are summarized in **Table 3,** and the corresponding confusion matrix and ROC curve are shown in **Fig. 6.**

Table 3: Final model performance on test set (t = 0.46)

Metric	ROC-AUC	Recall	Precision	F1-Score	Accuracy	Specificity	TP	FN	TN	FP
Value	0.9262	0.8785	0.5030	0.6397	0.8091	0.7927	253	35	955	250

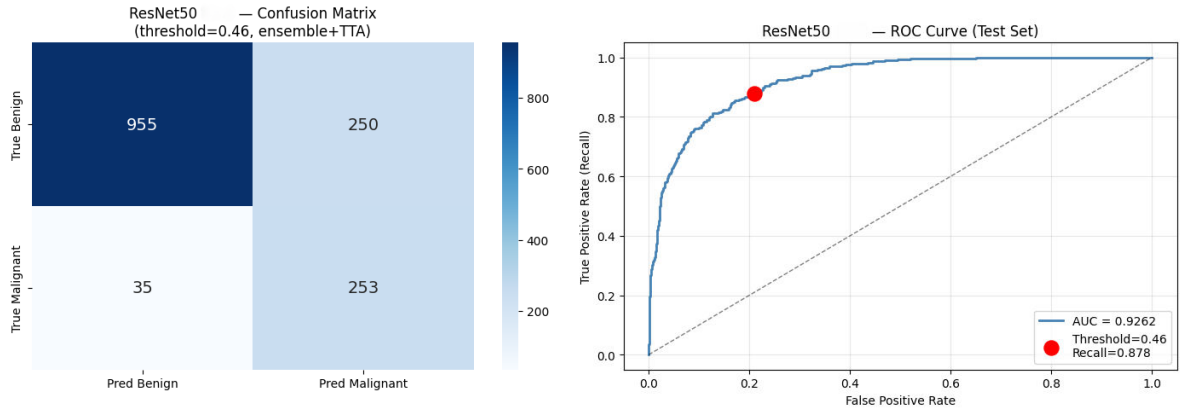


Fig. 6. Confusion matrix and ROC curve of the final ensemble model on the test set (t = 0.46).

The model achieves a ROC-AUC of 0.9262 and a recall of 0.8785, correctly identifying 253 out of 288 malignant cases. A total of 35 false negatives are observed, representing the most critical errors in a clinical context. These errors are primarily associated with atypical lesion appearance, imaging artifacts such as hair or markers, and low-contrast cases. The precision of 0.503 reflects the recall-oriented design of the system. While this results in a higher number of false positives, it aligns with clinical screening requirements where sensitivity is prioritized over specificity. Overall, the results indicate that the proposed framework achieves a strong balance between predictive performance and clinical relevance, with a clear focus on minimizing missed malignant cases.

Ablation Study

To evaluate the contribution of each component in the proposed pipeline, an ablation study was performed by incrementally adding components to a baseline ResNet50 model. The results are presented in **Table 4,** with performance trends illustrated in **Fig. 7.**

© 2026 IJMRSET | Volume 9, Issue 4, April 2026| DOI:10.15680/IJMRSET.2026.0904460

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Table 4: Ablation study — incremental component contribution

Configuration	AUC	Recall	Precision	F1	FN	FP
Baseline (single fold, t=0.5)	0.8915	0.7639	0.5450	0.6360	68	184
+ Differential LR + scheduler	0.9159	0.8194	0.5097	0.6285	52	227
+Label smoothing + class weights	0.9176	0.7670	0.6280	0.6910	67	131
+ Recall-optimized threshold	0.9176	0.8500	0.5400	0.6610	43	201
+ 3-fold CV	0.9220	0.8610	0.5100	0.6390	40	245
+ TTA ensemble (full pipeline)	**0.9262**	**0.8785**	**0.5030**	**0.6397**	**35**	**250**

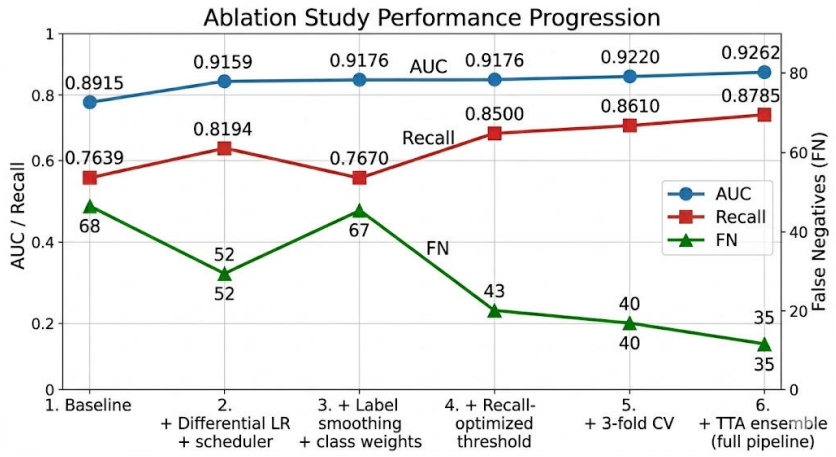


Fig. 7. Performance progression across ablation steps, showing improvements in AUC, recall, and reduction in false negatives.

The baseline model achieves a ROC-AUC of 0.8915 and a recall of 0.7639. Introducing differential learning rates and cosine annealing yields the most significant improvement, increasing AUC to 0.9159 and recall to 0.8194, highlighting the importance of effective fine-tuning. The addition of label smoothing and class weighting improves probability calibration and reduces false positives. Applying recall-oriented threshold optimization further reduces false negatives from 67 to 43 without altering model weights. Subsequent integration of 3-fold cross-validation and test-time augmentation leads to consistent performance gains, resulting in a final AUC of 0.9262 and reducing false negatives to 35. Overall, the complete pipeline reduces missed malignant cases from 68 to 35, demonstrating the cumulative benefit of each component.

Comparison with Existing Literature

To contextualize the proposed approach, **Table 5.** compares its performance with representative methods on the ISIC dataset.

Table 5: Comparison with published methods on ISIC dataset

Method	Backbone	AUC	Recall	Year
Esteva et al.	GoogLeNet	0.830	N/R	2017
Codella et al.	Ensemble CNN	0.911	0.730	2018
Kassani et al.	EfficientNet-B3	0.921	0.810	2020
Xie et al.	ResNet + transformer	0.924	N/R	2022
Kang et al.	DenseNet + augmentation	0.918	0.840	2022
This Work	**ResNet50 + ensemble**	**0.926**	**0.879**	**2026**

The model achieves a ROC-AUC of 0.926 and a recall of 0.879, demonstrating competitive performance. While several prior methods report similar AUC values, the proposed framework achieves higher recall, reflecting its emphasis on

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

minimizing false negatives. This highlights the advantage of a recall-oriented design in clinical settings, where missing malignant cases carries significant risk compared to false positives.

Qualitative Analysis — Grad-CAM++ Heatmaps

Representative Grad-CAM++ visualizations are shown in **Fig. 8** for true positive (TP), true negative (TN), false negative (FN), and false positive (FP) cases. Each example includes the original image, raw heatmap, and sharpened overlay.

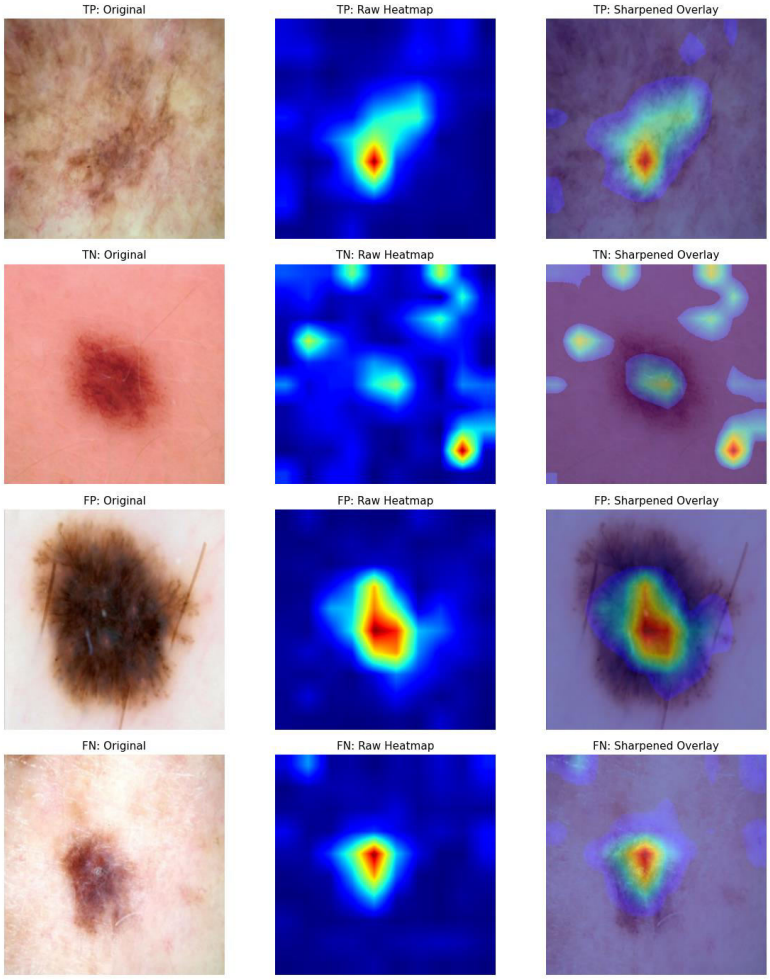


Fig. 8. Grad-CAM++ visualizations for TP, TN, FN, and FP cases. Columns show the original image, raw heatmap, and sharpened overlay.

For true positives, activation maps are concentrated around the lesion core and irregular borders, indicating attention to clinically relevant features. True negatives exhibit diffuse, low-intensity activations, suggesting the absence of strong malignant characteristics. False negatives reveal common failure patterns, including poor lesion localization, distraction by artifacts, and weak activation over lesion regions. False positives typically occur in benign lesions with atypical visual patterns, often near the decision boundary. Overall, the visualizations indicate that the model generally focuses on meaningful features, while errors are primarily associated with challenging lesion characteristics and external artifacts.

Error Analysis

The 35 false negative cases were systematically analyzed and categorized, as shown in **Table 6**.

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Table 6: False negative failure mode categorization

Failure Mode	Count	% of FN	Description
Artifact distraction	11	31.4%	Model attends to hair, stitches, or rulers rather than lesion
Atypical presentation	10	28.6%	Lesion lacks canonical malignancy visual markers
Poor lesion localization	8	22.9%	Activations scattered, lesion not identified
Low contrast (dark skin)	6	17.1%	Insufficient lesion–background contrast

Artifact distraction is the most common failure mode, where the model focuses on hair, rulers, or other non-diagnostic elements. Atypical presentation accounts for cases where malignant lesions lack clear visual indicators. Poor lesion localization occurs when activations are scattered and fail to align with the lesion region. Additionally, low-contrast cases, particularly in darker skin tones, show insufficient separation between lesion and background. These findings suggest that most errors are driven by visual complexity and dataset limitations rather than model instability. Improvements in preprocessing (e.g., artifact removal) and increased data diversity could further reduce false negatives.

Fairness Analysis

To assess performance across different skin tones, the test set was divided using the Individual Typology Angle (ITA) into light (ITA > 28) and dark (ITA ≤ 28) groups. The distribution is presented in **Table 7,** while performance metrics are summarized in **Table 8** and visualized in **Fig. 9.**

Table 7: Test set skin tone distribution

Group	ITA Range	N	Benign	Malignant	% of Test Set
Light skin	ITA > 28	864	721	143	57.9%
Dark skin	ITA ≤ 28	629	484	145	42.1%
Total	—	**1,493**	**1,205**	**288**	**100%**

Table 8: Model performance by skin tone at default threshold (t=0.46)

Metric	All	Light Skin	Dark Skin	Gap (L−D)	Significant?
ROC-AUC	0.8838	0.9087	0.8516	0.0571	Yes (>0.02)
Recall	0.7951	0.8112	0.7793	0.0319	No (<0.05)
Precision	0.5338	0.5472	0.5207	0.0265	No (<0.05)
F1-Score	0.6388	0.6535	0.6243	0.0292	No (<0.05)
FN (missed)	59	27	32	—	—
FN Rate	0.2049	0.1888	0.2207	0.0319	No (<0.05)

© 2026 IJMRSET | Volume 9, Issue 4, April 2026| DOI:10.15680/IJMRSET.2026.0904460

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

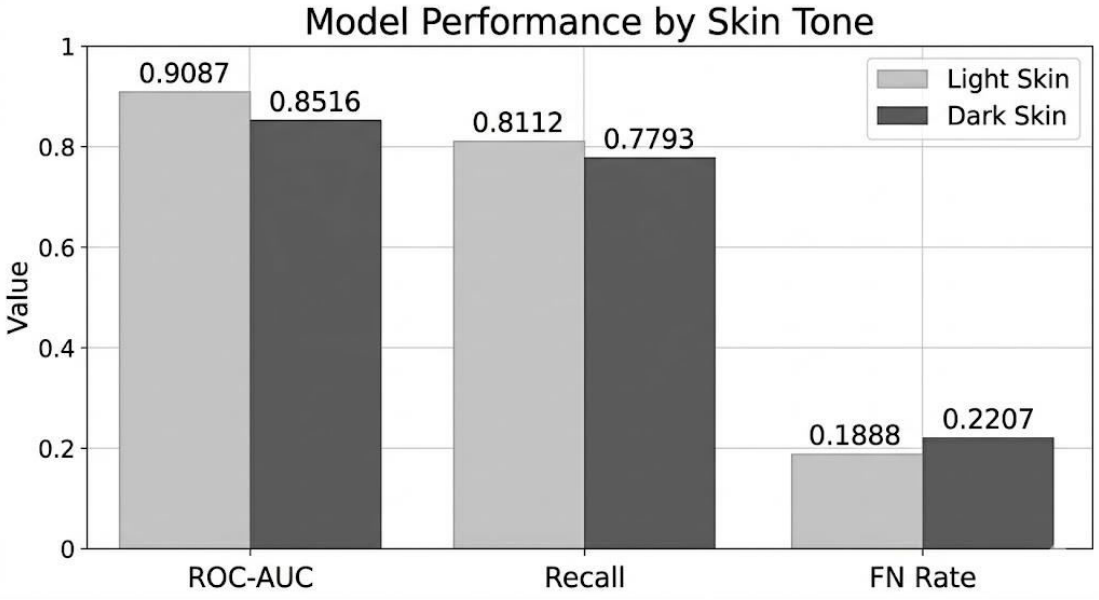


Fig. 9. Model performance comparison between light and dark skin groups across ROC-AUC, recall, and false negative rate.

A performance gap is observed between the two groups, with an AUC difference of 0.0571 (0.9087 vs. 0.8516), indicating reduced performance on darker skin images. Although recall differences are smaller, the false negative rate is higher for dark skin, indicating a greater risk of missed diagnoses. This disparity is attributed to both dataset imbalance and differences in probability calibration, where malignant probabilities for darker skin images tend to be lower. To address this issue, threshold recalibration is applied. The adjusted threshold ($t = 0.38$) improves recall across both groups and reduces the recall gap by approximately 42%, lowering total false negatives from 59 to 45. These results suggest that while some bias remains, a significant portion can be mitigated through threshold adjustment without retraining the model.

Limitations

Several limitations should be noted. First, all experiments are based on the ISIC dataset, and performance on other clinical datasets has not been fully validated. Domain shift may lead to reduced performance in real-world settings. Second, the ITA-based analysis relies on a proxy for skin tone rather than ground-truth Fitzpatrick annotations, which limits precision in fairness evaluation. Third, the Grad-CAM++ analysis is qualitative, as no quantitative localization metrics were used due to the lack of detailed segmentation annotations. Finally, the system has not been evaluated in a prospective clinical setting, and results should be interpreted as retrospective benchmark performance. Additionally, fairness analysis is limited to a binary skin tone grouping; a more detailed evaluation across multiple categories would provide deeper insights.

V. CONCLUSION

This study presents a deep learning-based framework for binary skin lesion classification, with a particular focus on reducing clinically critical errors such as false negatives. The proposed approach combines a ResNet50 backbone with transfer learning and a recall-oriented threshold optimization strategy to improve sensitivity toward malignant cases.

The final ensemble model, incorporating cross-validation and test-time augmentation, achieves a ROC-AUC of 0.926 and a recall of 0.879 on the ISIC dataset, demonstrating strong and reliable performance. Beyond predictive accuracy, the framework emphasizes interpretability and fairness. Grad-CAM++ is used to provide visual explanations of model predictions, indicating that the model consistently focuses on relevant lesion regions. Fairness analysis based on ITA highlights performance differences across skin tone groups, while threshold recalibration is shown to effectively reduce these disparities without requiring model retraining. Despite these promising results, certain limitations remain. Future

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

work will focus on improving dataset diversity to better address fairness, incorporating artifact removal techniques to reduce failure cases, and exploring more advanced architectures to enhance generalization. Additionally, evaluation in real-world clinical settings is necessary to assess the practical applicability of the proposed system.

REFERENCES

[1] American Cancer Society, "Cancer Facts & Figures 2023," 2023.
[2] A. Esteva et al., "Dermatologist-level classification of skin cancer with deep neural networks," Nature, vol. 542, no. 7639, pp. 115–118, 2017.
[3] N. Codella et al., "Skin lesion analysis toward melanoma detection," IEEE Journal of Biomedical and Health Informatics, vol. 23, no. 2, pp. 547–559, 2019.
[4] W. Stolz et al., "ABCD rule of dermatoscopy," The Lancet, vol. 349, pp. 1710–1711, 1997.
[5] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," Nature, vol. 521, pp. 436–444, 2015.
[6] K. He et al., "Deep residual learning for image recognition," in Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.
[7] R. R. Selvaraju et al., "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in Proc. IEEE International Conference on Computer Vision (ICCV), 2017.
[8] J. Gessert et al., "Skin lesion classification using ensembles of multi-resolution EfficientNets with meta data," Medical Image Analysis, 2020.
[9] A. Litjens et al., "A survey on deep learning in medical image analysis," Medical Image Analysis, vol. 42, pp. 60–88, 2017.
[10] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014.
[11] K. He et al., "ResNet: Deep residual networks," CVPR, 2016.
[12] G. Huang et al., "Densely connected convolutional networks," CVPR, 2017.
[13] A. Howard et al., "MobileNets: Efficient convolutional neural networks for mobile vision applications," 2017.
[14] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," ICML, 2019.
[15] S. Qin et al., "Hybrid CNN architectures for skin lesion classification," 2018.
[16] S. Rahman et al., "Deep learning-based multi-model framework for skin lesion classification," 2020.
[17] S. Hosny et al., "Skin cancer classification using transfer learning with AlexNet," 2020.
[18] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," Journal of Big Data, 2019.
[19] N. V. Chawla et al., "SMOTE: Synthetic minority over-sampling technique," Journal of Artificial Intelligence Research, 2002.
[20] E. D. Cubuk et al., "AutoAugment: Learning augmentation policies from data," CVPR, 2019.
[21] ISIC Archive, "International Skin Imaging Collaboration Dataset," https://www.isic-archive.com
[22] M. Kassani et al., "Skin lesion classification using deep learning," 2020.
[23] H. He and E. Garcia, "Learning from imbalanced data," IEEE Transactions on Knowledge and Data Engineering, 2009.
[24] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation," 1995.
[25] T. G. Dietterich, "Ensemble methods in machine learning," 2000.
[26] M. Rahman et al., "Skin lesion classification using ensemble learning techniques," 2020.
[27] P. Akilandasowmya et al., "Skin lesion classification using ResNet50," 2021.
[28] E. Del Bino et al., "The Individual Typology Angle (ITA): A skin color classification method," 2013.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY RESEARCH

IN SCIENCE, ENGINEERING AND TECHNOLOGY

| Mobile No: +91-6381907438 | Whatsapp: +91-6381907438 | ijmrset@gmail.com |

www.ijmrset.com