

A Hybrid U-Net and EfficientNet-B0 Framework for Skin Lesion Segmentation and Classification with Grad-CAM Explainability

Saad Rehman Lodhi ^a, Sobia Arshad ^b

^aPhD Student, HITEC University, Taxila, Pakistan

^bDepartment of Computer Engineering, HITEC University, Taxila, Pakistan

Received: 06 June 2026, Revised: 22 July 2026, Accepted: 29 September 2026, Published: 01 October 2026

KEY WORDS

Skin Lesion
Melasma
Dice-score
Segmentation
Classification
IoU

ABSTRACT

The aggressive nature of skin cancer requires timely and accurate diagnosis to improve patient outcomes. This study proposes an integrated deep learning framework that performs skin lesion segmentation and multi-class classification within a unified pipeline to streamline diagnosis and enhance predictive performance. The framework combines U-Net for lesion segmentation with EfficientNet-B0 for diagnostic classification, fusing dermoscopic image features with structured patient metadata (age, sex, and anatomical site). The segmentation model is evaluated on the ISIC 2018 dataset, whereas the classification model is trained and validated on the HAM10000 dataset. At a 0.55 threshold, the segmentation model achieved an Intersection over Union (IoU) of 0.9327 and a Dice score of 0.9652, demonstrating precise boundary delineation. The classification model achieved 86.0% accuracy, a macro AUC of 0.9773, and a macro F1-score of 0.7753, indicating strong multi-class generalization. Furthermore, Gradient-weighted Class Activation Mapping (Grad-CAM) visualizations improved model interpretability by highlighting the spatial regions driving diagnostic predictions, particularly in misclassified cases.

1. Introduction

Automated skin lesion analysis relies on two core pillars: advanced deep learning architectures and curated dermatological imaging datasets. Many contemporary approaches exhibit limited diagnostic accuracy and contextual awareness because they treat lesion segmentation and classification as isolated tasks [1]. In medical image analysis, convolutional neural networks (CNNs) and attention-based architectures—most notably U-Net for segmentation and EfficientNet for classification—have demonstrated strong individual performance [1]. However, integrating visual explainability with pixel-level ground-truth supervision remains challenging in real-world clinical settings, where transparent decision rationales are essential for clinical adoption [2]. This study proposes a unified deep learning framework that performs skin lesion segmentation and multi-class classification within a single pipeline, incorporating both clinical metadata fusion and Grad-CAM explainability. Furthermore, most existing models neglect clinical metadata such as lesion localization, patient sex, and age, despite their ability to improve diagnostic discrimination [3,4]. Accurate, timely, and explainable skin cancer diagnosis directly informs clinical treatment pathways and patient outcomes. Providing visual rationales via Grad-CAM aligns automated predictions with dermatologists' diagnostic reasoning, strengthening clinical trust

and usability. The proposed framework evaluates seven major skin lesion categories and fuses localization, sex, and age metadata [5]. Although the framework has not yet been tested in prospective clinical trials, it is empirically evaluated on benchmark datasets to support academic and pre-clinical validation [6].

1.1 Research Novelty and Contributions

Unlike conventional single-task approaches, this research integrates skin lesion segmentation, clinical metadata fusion, multi-class classification, and visual explainability within a unified framework. Specifically, the pipeline combines U-Net-based lesion segmentation with an EfficientNet-B0 classification network that incorporates patient sex, age, and anatomical site, while employing Grad-CAM to interpret classification decisions and strengthen clinical transparency.

The primary contributions of this research are as follows:

1. Developing an end-to-end framework that links skin lesion segmentation and multi-class classification.
2. Integrating clinical metadata (age, sex, and anatomical site) with dermoscopic image features to improve diagnostic performance.
3. Incorporating Grad-CAM visualizations to enhance model interpretability and clinical transparency.
4. Conducting a comprehensive evaluation using classification metrics (accuracy, macro F1-score, and

macro AUC) and segmentation metrics (IoU and Dice score).

5. Demonstrating the feasibility of explainable multi-modal AI systems for dermatological diagnosis.

2. Literature Review

This section reviews relevant literature on automated skin lesion analysis, spanning classical image processing, deep learning models for segmentation and classification, unified multi-task architectures, and explainability in medical imaging.

2.1 Related Work

2.1.1 Traditional and Early Approaches

Conventional skin lesion analysis relied on classical image processing and machine learning algorithms, including k-Nearest Neighbors (k-NN), Random Forests, and Support Vector Machines (SVMs) [7], operating on handcrafted shape features, texture descriptors (GLCM and LBP), and color histograms [8]. While effective under controlled conditions, these methods degrade in the presence of illumination variations, hair artifacts, noise, and irregular boundaries, limiting their generalizability and motivating deep-learning-based solutions.

2.1.2 Deep Learning for Skin Lesion Analysis

Segmentation and classification accuracy improve substantially when deep Convolutional Neural Networks (CNNs) are employed. U-Net and its variants serve as the primary architecture for dermoscopic segmentation due to their ability to capture multi-scale spatial context and preserve fine-grained boundary details. However, standalone segmentation models remain sensitive to annotation quality, imaging artifacts, and class imbalance. For classification, deep CNN architectures such as EfficientNet, DenseNet, and ResNet learn hierarchical representations directly from pixel data [9], with EfficientNet balancing computational efficiency and accuracy through compound scaling. Nevertheless, most classification models rely solely on image data and overlook clinical metadata—such as anatomical site, sex, and age—that are integral to dermatological decision-making [10]. Moreover, treating segmentation and classification as separate pipelines can introduce inconsistencies in feature learning and error propagation, where inaccurate segmentation degrades downstream classification.

2.1.3 Multi-Task Learning and Unified Frameworks

To overcome the limitations of separate pipelines, multi-task learning (MTL) and unified frameworks combine segmentation and classification through shared feature extraction and task-specific branches, improving data efficiency and robustness. However, unified frameworks face challenges including increased architectural complexity, task imbalance, and negative transfer between tasks [11]. Furthermore, few unified pipelines

incorporate structured clinical metadata alongside image features, limiting their ability to model real-world diagnostic workflows.

2.1.4 Explainability in Medical Imaging

To deploy deep learning models effectively in clinical settings, explainability has become a core requirement. Visual explanation methods such as Gradient-weighted Class Activation Mapping (Grad-CAM) highlight the image regions that most strongly influence model predictions [12], making diagnostic decisions more transparent and aligned with clinical reasoning. However, the sensitivity of Grad-CAM visualizations depends on the underlying network architecture, which can limit reliability in complex diagnostic cases [13]. Table 1 summarizes the key research gaps in existing literature and how they are addressed in this study.

2.2 Abbreviation Table

3. Proposed Methodology

This section presents the proposed methodology across dataset selection, preprocessing, model architecture, training, and evaluation (Fig. 1). To address the challenges of dermoscopic image analysis, the framework combines CNN-based U-Net segmentation, multimodal EfficientNet-B0 classification, and Grad-CAM visual explainability.

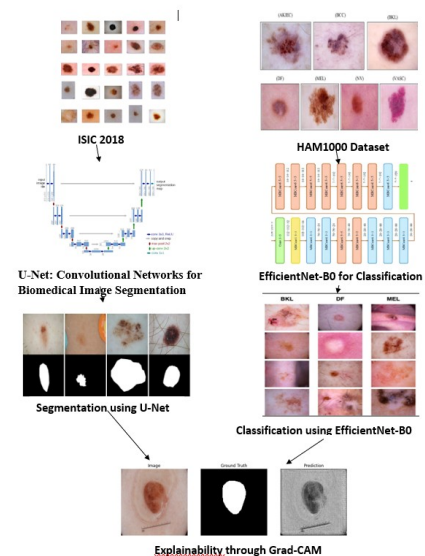


Fig. 1. Workflow Overview

3.1 Research Design

This research adopts an applied, experimental methodology within a quantitative evaluation framework. A two-stage deep learning pipeline was developed comprising (1) a U-Net architecture for binary lesion segmentation and (2) an EfficientNet-B0 network for seven-class lesion classification with metadata fusion. Model performance was evaluated using standard metrics: Intersection over Union (IoU) and Dice coefficient for segmentation, and accuracy, macro F1-score, and macro AUC for classification.

Table 1
Research Gaps and Contributions

Research Gap	Description	Addressed by This Study
1. Lack of unified models.	Most studies perform segmentation and classification as two separate tasks [14].	Proposed a joint segmentation-classification pipeline with shared preprocessing and lesion-focused cropping.
2. Limited interpretability in deep learning models.	Many black-box models lack transparency, making clinical adoption challenging.	Integrated Grad-CAM for visual explanation of classification decisions to improve interpretability.
3. Class imbalance in dermatology datasets.	Lesion datasets like HAM10000 are biased toward common classes (e.g., NV), limiting rare class detection.	Applied augmentation and metadata fusion (age, sex, site) to enhance feature learning for all classes.
4. Poor generalization across lesion types and metadata variation.	Few models leverage clinical metadata alongside images.	Combined image data with structured metadata during classification using EfficientNet-B0.
5. Overreliance on supervised data for segmentation.	Manual mask annotation is expensive; reliance on fully labelled data limits scalability.	While the current model is supervised, future directions suggest semi-supervised learning with pseudo-labelling.
6. Lack of integrated evaluation and visualization tools.	Evaluation is often metric-centric without visual feedback.	Combine quantitative metrics (Dice, AUC, F1) with qualitative outputs (segmentation masks, Grad-CAMs).

Table 2
Abbreviation Table

Abbreviation	Full Form	Abbreviation	Full Form
AI	Artificial Intelligence	CRF	Conditional Random Field
AUC	Area under the ROC Curve	DF	Dermatofibroma
BCE	Binary Cross-Entropy	DL	Deep Learning
BCC	Basal Cell Carcinoma	ECC	EfficientNet Classification Component
BKL	Benign Keratosis-Like Lesion	F1-Score	Harmonic Mean of Precision and Recall
CAM	Class Activation Mapping	FN	False Negative
CNN	Convolutional Neural Network	FP	False Positive
GAM	Global Average Max-Pooling	MEL	Melanoma
GLCM	Gray Level Co-occurrence Matrix	ML	Machine Learning
Grad-CAM	Gradient-Weighted Class Activation Mapping	MLP	Multi-Layer Perceptron
HAM10000	Human Against Machine 10000 Dataset	MTL	Multi-Task Learning
IoU	Intersection over Union (Jaccard Index)	NV	Melanocytic Nevus
ISIC	International Skin Imaging Collaboration	ReLU	Rectified Linear Unit
KNN/k-NN	k-Nearest Neighbors	RF	Random Forest
LBP	Local Binary Patterns	ROC	Receiver Operating Characteristics
SVM	Support Vector Machine	U-Net	Convolutional Encoder-Decoder Model for Segmentation
SSL	Semi-Supervised Learning	ViT	Vision Transformer
TN	True Negative	VASC	Vascular Lesion
TP	True Positive	XAI	Explainable Artificial Intelligence

3.2 Dataset and Data Acquisition

The HAM10000 (Human Against Machine 10,000) dataset, retrieved from Kaggle [15], serves as the primary data source for the classification task. It comprises 10,015 dermoscopic images categorized into seven diagnostic classes, as summarized in Table 3 and illustrated in Fig. 2. During data preparation, image files from the source directories were consolidated and linked to the structured metadata file (HAM10000_metadata.csv), which maps each image identifier to its diagnosis label,

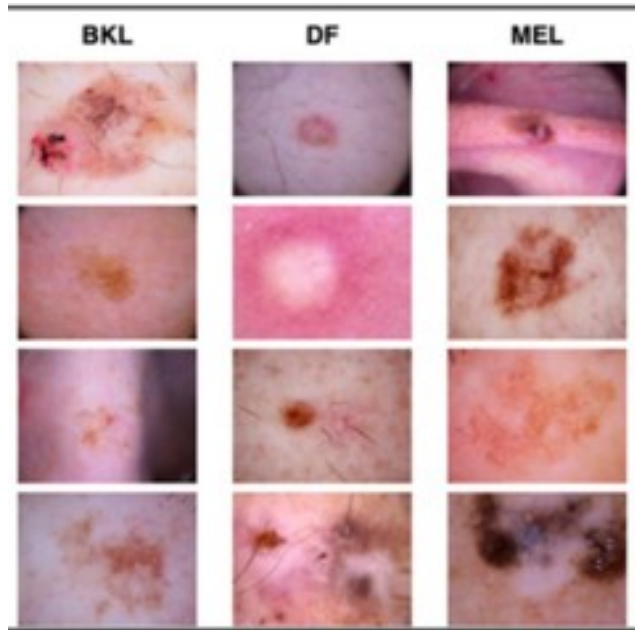
file path, and patient attributes (age, sex, and anatomical site).

The structured metadata attributes (anatomical site, sex, and age) were merged with the corresponding image paths using unique image IDs. By combining dermoscopic images, clinical metadata, and expert-verified labels, HAM10000 serves as a comprehensive benchmark for developing and evaluating skin lesion classification models [16]. For empirical evaluation, 80% of the dataset was used for training and 20% was reserved for validation.

Table 3

Classes of the HAM10000 Dataset

Class	Full Name	Images
NV	Melanocytic Nevus	6,705
MEL	Melanoma	1,113
BKL	Benign Keratosis-Like Lesion	1,099
BCC	Basal Cell Carcinoma	514
AKIEC	Actinic Keratosis / Bowen's Disease	327
DF	Dermatofibroma	115
VASC	Vascular Lesion	142

**Fig. 2.** Sample Dermoscopic Images from the HAM10000 Dataset

For the segmentation task, the ISIC 2018 Skin Lesion Segmentation Challenge dataset was used. The dataset comprises 2,594 dermoscopic images, each paired with an expert-annotated binary lesion mask that provides pixel-level supervision delineating the lesion from surrounding skin.

This dataset is essential to:

- Evaluate segmentation generalization across diverse lesion morphologies.
- Assess boundary delineation on irregular and asymmetric lesion borders.
- Test boundary sensitivity under low-contrast skin conditions.

Segmentation in ISIC 2018 is formulated as a binary task (lesion versus background), with images spanning multiple diagnostic categories [17], as summarized for the classification subset in Table 4.

Although segmentation labels are binary, the underlying images span diverse diagnostic classes, helping the segmentation model accommodate variations in lesion color, texture, and shape. Across the 2,594-image ISIC 2018 segmentation dataset, each sample includes:

- Expert dermatological annotations of lesion boundaries.

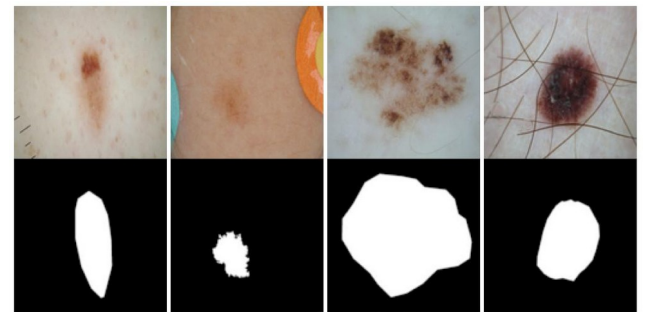
Table 4

Classes of the ISIC 2018 Dataset

Class	Full Name	Number of Images
NV	Melanocytic Nevus	1,262
MEL	Melanoma	374
BKL	Benign Keratosis-Like Lesions	262
BCC	Basal Cell Carcinoma	332
AKIEC	Actinic Keratosis / Bowen's Disease	65
DF	Dermatofibroma	43
VASC	Vascular Lesion	26
Total Images	—	2,364 (classification subset)

- Consistent spatial alignment between dermoscopic images and masks.
- A corresponding expert-drawn binary segmentation mask (.png).
- A high-resolution dermoscopic lesion image (.jpg).

Fig. 3 illustrates representative dermoscopic images paired with their expert-annotated binary masks, enabling pixel-wise comparison between predicted masks and ground truth. For segmentation, the dataset was split into 70% for training, 20% for validation (hyperparameter tuning), and 10% for testing.

**Fig. 3.** Sample Images and Ground-Truth Binary Masks from the ISIC 2018 Dataset

3.3 Data Preprocessing and Augmentation

Both the HAM10000 and ISIC 2018 datasets underwent standardized preprocessing. All images and binary masks were resized to 224×224 pixels, normalized, and spatially aligned. Data augmentation—including random rotations, horizontal and vertical flips, and color-space adjustments—was applied during training, while HAM10000 images were linked to encoded metadata attributes (anatomical site, age, and sex).

Representative dermoscopic images after data augmentation are shown in Fig. 4. Data augmentation is critical for improving model generalization; incorporating rotations, flips, and color-space perturbations mitigates class imbalance and helps the model learn invariant lesion features [18].

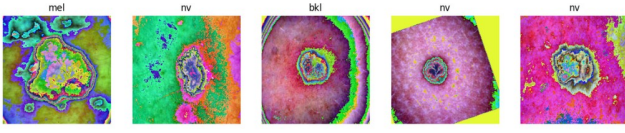


Fig. 4. Augmented Training Samples

3.4 Model Architecture

The segmentation module is based on the U-Net architecture implemented in PyTorch [19], whereas the classification module employs EfficientNet-B0 pre-trained on ImageNet. The final fully connected layer of EfficientNet-B0 was modified to output class probabilities across the seven skin lesion categories. The network was fine-tuned using the Adam optimizer and cross-entropy loss [20] with a batch size of 32 over 5 epochs across the training and validation splits. EfficientNet-B0 was selected for its compound scaling strategy, which balances efficiency and accuracy while outperforming heavier CNNs such as ResNet with reduced computational load and overfitting [21]. Additionally, EfficientNet-B0 supports metadata fusion by concatenating encoded metadata with image features prior to final prediction, improving diagnostic accuracy on complex lesions.

3.5 Unified Multi-Task Framework

Integrating segmentation and classification within a unified pipeline enables joint learning of pixel-level boundary localization and image-level diagnostic features. Binary Cross-Entropy (BCE) loss was used to train the segmentation model alongside cross-entropy loss for classification [22], with training accelerated via GPU (CUDA). Grad-CAM was integrated to interpret model predictions—particularly misclassified cases—alongside confusion matrices. Because Grad-CAM is class-discriminative and architecture-agnostic, it provides visual explanations without modifying the underlying network.

3.6 Evaluation Metrics

To evaluate the unified segmentation and classification framework, task-specific quantitative metrics were employed.

3.6.1 Segmentation Evaluation Metrics

The Dice coefficient and Intersection over Union (IoU) were used to measure spatial overlap between ground-truth lesion areas and predicted binary masks [23].

3.6.1.1. Dice Coefficient

The Dice coefficient, also known as the Sørensen–Dice index [24], measures spatial similarity between the ground-truth mask and the predicted mask and is particularly sensitive to small lesion structures:

$$Dice = \frac{2 \cdot |A \cap B|}{|A| + |B|}$$

where A denotes the predicted binary mask and B denotes the ground-truth binary mask.

Precise boundary delineation directly improves lesion-focused cropping prior to downstream classification.

3.6.1.2. Intersection over Union (IoU) / Jaccard Index

IoU measures the spatial overlap between the ground-truth and predicted lesion areas divided by their union [23, 24]:

$$IoU = \frac{|A \cap B|}{|A \cup B|}$$

IoU provides a rigorous measure of region-based segmentation quality by penalizing both over-segmentation and under-segmentation.

3.6.2 Classification Evaluation Metrics

Classification performance across the seven lesion categories was evaluated using the following metrics:

3.6.2.1. Accuracy

Measures the overall proportion of correctly classified samples:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

While accuracy provides a concise summary of overall performance, it can be skewed by majority classes in imbalanced datasets.

3.6.2.2. Macro F1-Score

Computes the unweighted average of the F1-scores calculated independently across all $N = 7$ classes:

$$F1 = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall}$$

$$Macro F1 = \frac{1}{N} \sum_{i=1}^N F1_i$$

Macro F1-score is well suited for imbalanced datasets such as HAM10000, ensuring that rare classes (e.g., VASC and DF) are weighted equally alongside common classes.

3.6.2.3. Macro AUC

Evaluates the model's ability to discriminate between lesion classes across all decision thresholds [25]. Threshold-independent AUC values quantify the classifier's ranking capability, which is critical in medical diagnosis.

3.6.2.4. Confusion Matrix

A cross-tabulation of predicted versus true classes that reveals class-specific misclassification patterns and areas for improvement.

3.6.2.5. Per-Class ROC Curves

Plots the true positive rate against the false positive rate for each class, providing class-wise insight into diagnostic separability.

3.7 Explainability with Grad-CAM

Gradient-weighted Class Activation Mapping (Grad-CAM) generates activation heatmaps that highlight the spatial regions most influential to model decisions [26], diagnosing misclassifications, enhancing transparency, and building clinical trust.

4. Results and Discussion

This section presents and analyzes the experimental results for both lesion segmentation and multi-class classification, detailing the hardware and software environment, model selection rationale, threshold sensitivity analysis, and quantitative and qualitative performance.

4.1 Hardware and Software Specifications

Experiments were conducted in Google Colaboratory using an NVIDIA Tesla T4 GPU (16 GB VRAM) detected via CUDA and a virtual multicore CPU (12–25 GB RAM). Google Drive was used to store checkpoints, training logs, and datasets to ensure reproducibility across sessions. Model development was implemented in Python and PyTorch, using the segmentation-models-pytorch (SMP) library to instantiate the U-Net architecture with an ImageNet-pretrained ResNet-34 encoder.

NumPy, OpenCV, and the Python Imaging Library (PIL) handled image reading and mask transformations, while Matplotlib and scikit-learn supported visualization and dataset partitioning. PyTorch's built-in CUDA/cuDNN integration and automated dependency management in Colaboratory (including tqdm, torchvision, and SMP), together with PyTorch data loaders and tensor pipelines, provided full control over preprocessing, normalization, batching, and augmentation.

4.2 Segmentation Results

Fig. 5 summarizes the segmentation model's training progression across epochs, reporting the training loss alongside the Dice score and IoU. Each successive epoch reduced loss while increasing both IoU and Dice scores, demonstrating steady feature learning. By the fourth and fifth epochs, the IoU and Dice metrics stabilized (Epoch 5: Loss = 0.2151, Dice = 0.8816, IoU = 0.8058), indicating convergence. This inverse relationship between loss reduction and spatial overlap confirms that the U-Net model effectively learned the structural and spatial features of skin lesions.

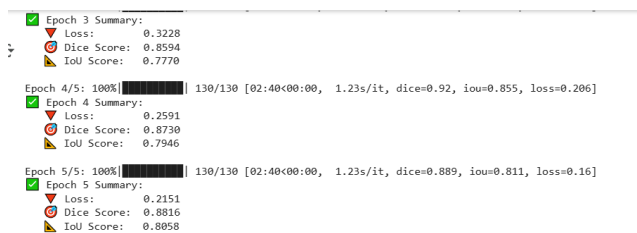


Fig. 5. Segmentation Training Progression Across Epochs 3–5

Visual inspection of the segmentation outputs corroborates the quantitative metrics. Fig. 6 shows a repre-

sentative validation sample, demonstrating close agreement between the predicted binary mask and the expert ground truth across the ISIC 2018 dataset.

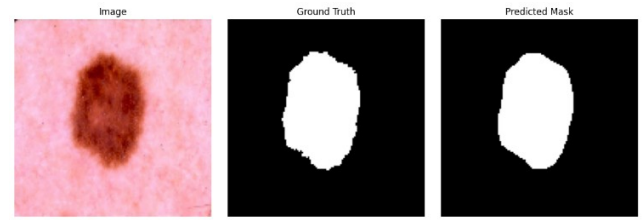


Fig. 6. Validation Segmentation Sample (Original Image, Ground Truth, and Predicted Mask)

4.2.1 Threshold Tuning and Post-Processing

The segmentation model outputs continuous probability maps in the range [0, 1] [27], representing the estimated probability that each pixel belongs to a lesion. Thresholding is required to convert these probability maps into binary masks. A threshold sensitivity analysis was conducted across thresholds from 0.30 to 0.70, as shown in Fig. 7.

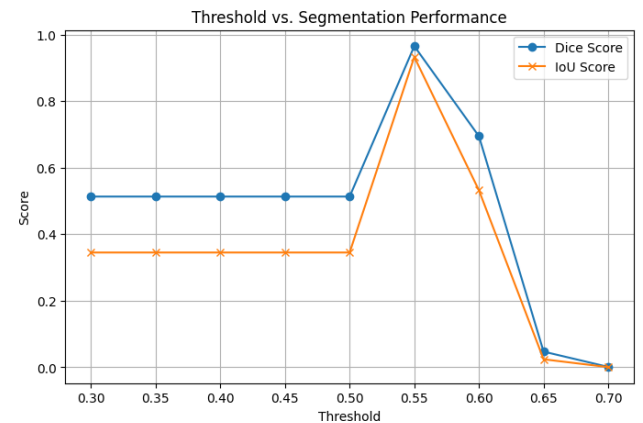


Fig. 7. Threshold Sensitivity Analysis vs. Segmentation Performance

Segmentation performance peaked at a threshold of 0.55, achieving an IoU score of 0.9327 and a Dice score of 0.9652 on well-defined lesion samples. However, a standard threshold of 0.50 was retained for general inference, as it provides more stable and generalizable predictions across lesions of varying contrast, size, and border irregularity while balancing false positives and false negatives. Above 0.60, segmentation performance drops sharply due to overly conservative thresholding that discards valid lesion regions. Representative three-panel comparisons of original images, ground-truth masks, and predicted outputs are presented in Fig. 8.

4.3 Classification Results

4.3.1 Classification Model and Training Performance

For multi-class skin lesion classification, EfficientNet-B0 was trained on the HAM10000 dataset across seven diagnostic categories, fusing dermoscopic image features with patient metadata (age, sex, and anatomical site) to balance computational efficiency and diagnostic accuracy.

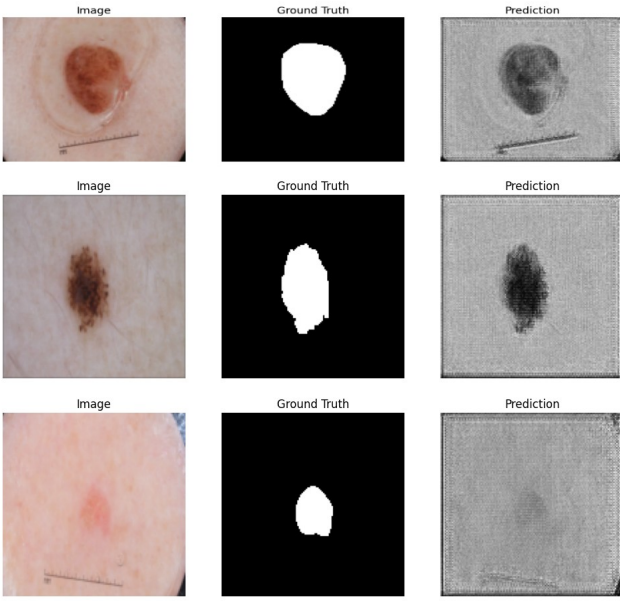


Fig. 8. Representative Segmentation Predictions Across Lesion Types

4.3.2 Rationale for Selecting EfficientNet-B0

EfficientNet-B0 leverages a compound scaling method that systematically balances network resolution, depth, and width. Compared with traditional CNNs such as Inception or ResNet, EfficientNet-B0 achieves higher accuracy with substantially fewer parameters [28], mitigating overfitting and data-scarcity challenges in medical imaging.

4.3.3 Epoch-Wise Training Dynamics

Training was conducted for 5 epochs, tracking training and validation loss and accuracy across epochs, as illustrated in Fig. 9.

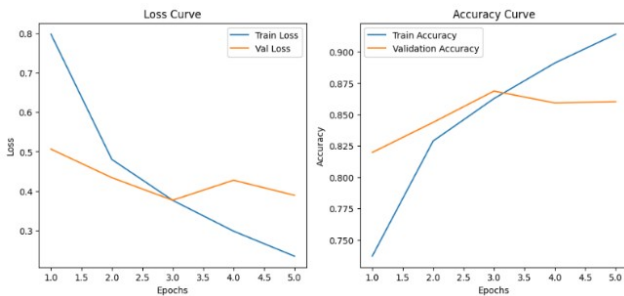


Fig. 9. Training and Validation Loss and Accuracy Curves

As shown in Fig. 9, both training loss and validation loss decreased steadily across epochs, indicating effective generalization and minimal overfitting. Validation accuracy stabilized after the third epoch and plateaued at 86.0%, confirming the model's learning capability.

4.3.4 Final Performance Metrics

The classification model was evaluated on the HAM10000 validation split using standard per-class classification metrics, as summarized in Table 5.

4.3.4.1. Interpretation of Metrics

Accuracy: Measures overall classification correctness; at 86.0% accuracy, approximately six out of every seven validation samples were classified correctly.

Table 5

Per-Class Classification Evaluation Metrics

Class	Precision	Recall	F1-score	Support
AKIEC	0.60	0.71	0.65	65
BCC	0.91	0.73	0.81	103
BKL	0.70	0.74	0.72	220
DF	0.75	0.78	0.77	23
MEL	0.65	0.68	0.68	223
NV	0.94	0.93	0.93	1341
VASC	0.89	0.87	0.87	28

Macro F1-Score: Represents the unweighted average F1-score across all classes. The achieved score of 0.7753 (~78%) indicates a robust balance between precision and recall across both common and rare lesion classes.

ROC Curves by Class: The per-class ROC curves evaluate discrimination between malignant and benign lesions, where an AUC of 1.0 denotes ideal separability. The achieved macro AUC of 0.9773 confirms high separability across class distributions (Fig. 10).

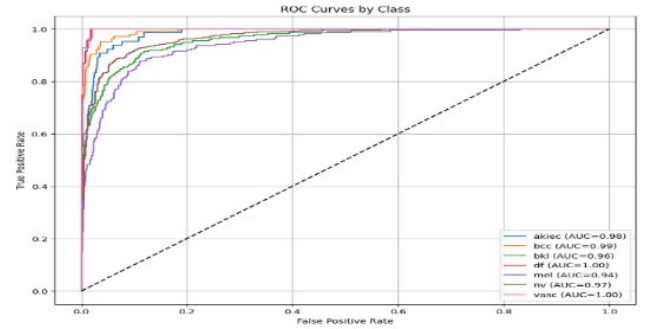


Fig. 10. Per-Class Receiver Operating Characteristic (ROC) Curves

4.3.5 Effect of Metadata Integration

A key source of the model's discriminative strength is the integration of structured clinical metadata:

- Age provides clinical context, as specific lesion categories exhibit distinct age-related prevalence.
- Sex captures demographic variation influencing lesion distribution and risk.
- Anatomical site encodes location-specific priors (e.g., torso vs. acral or scalp lesions), as shown in Table 6.

By fusing clinical metadata with image-based features (shape, texture, and color), the model learns more robust class boundaries, helping differentiate visually similar lesions such as melanocytic nevus (NV) and melanoma (MEL). This multimodal fusion strategy compensates for visual overlap, boosts classification sensitivity on underrepresented classes, and improves generalization to real-world clinical scenarios.

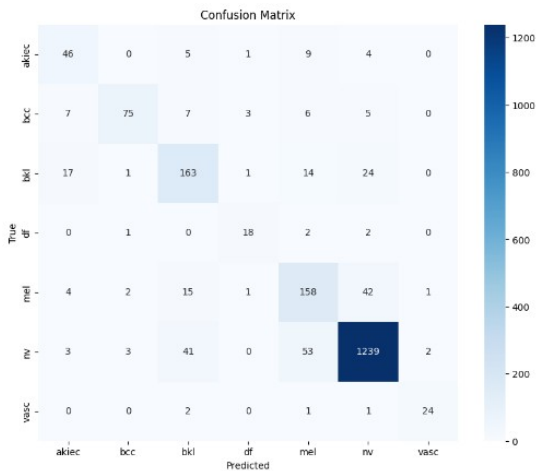
4.3.6 Confusion Matrix Analysis

The confusion matrix in Fig. 11 illustrates class-wise performance on the HAM10000 validation split. Performance on AKIEC is moderate, reflecting visual

Table 6

Encoded Clinical Metadata Records (HAM10000)

#	lesion_id	image_id	dx	dx_type	age	sex	localization	image_path
0	HAM_0000118	ISIC_0027419	bkl	histo	80.0	male	scalp	/content/ham10000/ final_images/ISIC_0027419.jpg
1	HAM_0000118	ISIC_0025030	bkl	histo	80.0	male	scalp	/content/ham10000/ final_images/ISIC_0025030.jpg
2	HAM_0002730	ISIC_0026769	bkl	histo	80.0	male	scalp	/content/ham10000/ final_images/ISIC_0026769.jpg
3	HAM_0002730	ISIC_0025661	bkl	histo	80.0	male	scalp	/content/ham10000/ final_images/ISIC_0025661.jpg
4	HAM_0001466	ISIC_0031633	bkl	histo	75.0	male	ear	/content/ham10000/ final_images/ISIC_0031633.jpg

**Fig. 11.** Multi-Class Confusion Matrix on the HAM10000 Validation Split

overlap with benign keratosis-like lesions (BKL) and nevus (NV) in color and texture. Strong performance is achieved for BCC, BKL, and especially NV, which benefits from distinct dermoscopic patterns and ample training representation. Notably, both dermatofibroma (DF) and vascular lesions (VASC) achieved high recall despite being minority classes, underscoring the benefit of metadata fusion [29].

Overall, the confusion matrix over the 20% validation split reflects the underlying class distribution of HAM10000. High diagonal counts for NV, BKL, MEL, and BCC confirm strong accuracy on dominant categories, while remaining misclassifications in rarer classes (AKIEC, VASC, and DF) stem from limited training samples.

4.4 Mapping of Segmentation and Classification Performance Metrics

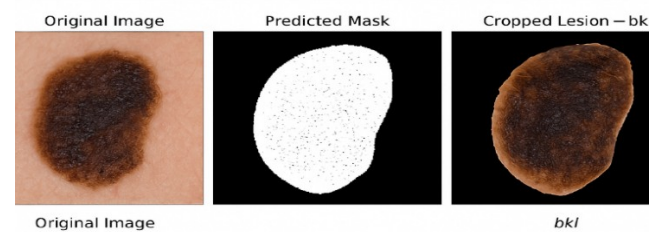
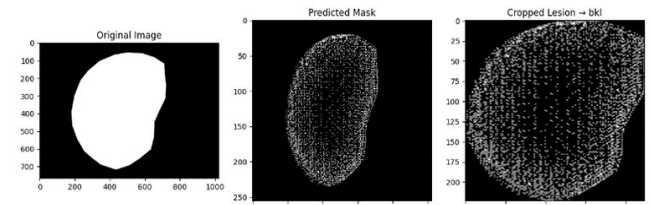
Although segmentation and classification operate within a unified pipeline, they generate distinct output representations and are therefore evaluated using task-specific performance metrics.

For segmentation, the model produces pixel-level binary masks that are compared pixel-by-pixel against expert ground-truth masks using IoU, Dice score, pixel accuracy, sensitivity, and specificity. IoU and Dice directly quantify spatial overlap and boundary delineation quality.

For classification, the model predicts a single diagnostic label per lesion image, which is evaluated at the image level using accuracy, precision, recall, macro F1-score, macro AUC-ROC, and the confusion matrix to measure diagnostic discrimination across categories (e.g., BCC, NV, and MEL).

The achieved segmentation scores (IoU = 0.9327, Dice = 0.9652) confirm precise boundary delineation, while the classification metrics (accuracy = 86.0%, macro F1-score = 0.7753, macro AUC = 0.9773) demonstrate strong multi-class discrimination despite severe class imbalance. Clinically, minimizing false negatives is vital to prevent diagnostic delays in skin cancer; these quantitative results highlight the framework's potential to assist dermatologists in early detection.

4.5 Grad-CAM Visualization and Model Explainability

**Fig. 12.** Original Dermoscopic Image, Predicted Mask, and Cropped Lesion (BKL)**Fig. 13.** Ground-Truth Binary Mask, Predicted Activation Map, and Cropped Lesion Visualization

Figs. 12 and 13 illustrate how the unified framework processes a representative sample from the ISIC 2018 dataset through segmentation, lesion-focused cropping, and classification. Fig. 12 shows the original dermoscopic image (left), the predicted binary mask (center), and the cropped lesion isolated for classification (right). Fig. 13 compares the clean ground-truth binary mask (left) with the predicted spatial activation map (center) and the resulting cropped lesion (right). Although the

raw predicted map exhibits minor boundary irregularity and speckle noise, it retains the structural fidelity required to isolate the lesion effectively during cropping. This demonstrates the robustness of the unified pipeline, where minor segmentation artifacts do not degrade downstream classification performance. The EfficientNet-B0 classifier correctly identifies the lesion as BKL (benign keratosis-like lesion), confirming strong end-to-end generalization.

4.6 Final Performance Comparison

Comparison against recent state-of-the-art studies (Table 7) demonstrates that the proposed framework achieves competitive segmentation performance (Dice = 0.9652) and classification accuracy (86.0%). Combining U-Net segmentation, metadata-fused EfficientNet-B0 classification, and Grad-CAM explainability strengthens both predictive performance and clinical transparency.

4.7 Practical Implications for Dermatological Diagnosis

The quantitative and qualitative findings underscore the framework's relevance for dermatological screening and clinical decision support. Accurate lesion segmentation assists clinicians in delineating lesion margins and tracking morphological progression, while automated classification helps reduce diagnostic workload. Integrating patient metadata alongside Grad-CAM visual explanations provides transparent, context-aware predictions that support early skin cancer detection and timely clinical intervention.

5. Conclusion and Future Work

5.1 Conclusion

In this study, a unified deep learning framework was implemented for automated skin lesion analysis, integrating U-Net segmentation (on ISIC 2018), EfficientNet-B0 classification with clinical metadata fusion (on HAM10000), and Grad-CAM visual explainability. In initial evaluation, the segmentation module achieved a Dice score of 0.6789, an IoU of 0.5544, and a pixel accuracy of 87.0%; following threshold tuning, segmentation performance peaked at an IoU of 0.9327 and a Dice score of 0.9652 at a 0.55 threshold, while a 0.50 threshold was retained for balanced generalization across unseen data.

On the classification side, the metadata-fused EfficientNet-B0 model achieved a final accuracy of 86.02%, a macro F1-score of 0.7753, and a macro AUC of 0.9773. Per-class evaluations confirmed strong generalization for prevalent lesion types such as melanocytic nevus (NV) and benign keratosis-like lesions (BKL), while metadata inclusion helped mitigate misclassifications on rare categories such as vascular lesions (VASC) and dermatofibroma (DF). Visual explanations via Grad-CAM localized the discriminative regions

influencing model predictions, supporting post-hoc validation and identifying failure modes where attention shifted to peripheral features. Overall, the integrated segmentation–classification pipeline improves both diagnostic accuracy and model transparency in medical imaging.

5.2 Limitations

Despite encouraging results, several limitations were observed. First, both the ISIC 2018 and HAM10000 datasets exhibit substantial class imbalance, with NV heavily overrepresented relative to VASC and DF, which may constrain sensitivity on rare lesion types. Second, the study relies on public benchmark datasets that may not fully reflect the diversity of clinical populations, acquisition conditions, skin tones, and imaging devices encountered in clinical practice; external validation on multi-institutional cohorts is therefore essential prior to deployment. Third, the segmentation model can produce noisy boundaries on low-contrast lesions with ambiguous edges; although these imperfections did not significantly degrade classification accuracy, they may affect boundary reliability in challenging clinical cases.

5.3 Future Work

Future research may explore hybrid CNN–Transformer architectures and Vision Transformers (ViTs) to improve long-range contextual modeling and feature extraction. Self-supervised and semi-supervised learning approaches with pseudo-labeling can also reduce reliance on manually annotated segmentation masks by leveraging unlabeled dermoscopic images. Furthermore, multimodal fusion can be extended to incorporate electronic health records, clinical photographs, and patient history, while complementary explainability methods—such as Score-CAM, LIME, and SHAP—and prospective clinical validation studies with dermatologists will further assess the framework's effectiveness in clinical practice.

6. References

- [1] A. Abdel-Wahab, V. Pulipati, S. Joshi, R. Srikanthswara, S. Menon, and P. Dutta, "Deep Learning Framework for Breast Cancer Detection and Segmentation Using EfficientNet and U-Net with Hyperparameter Optimization," *SN Computer Science*, vol. 6, no. 5, p. 528, 2025.
- [2] I. Matas, C. Serrano, F. Silva, A. Serrano, T. Toledo-Pastrana, and B. Acha, "Clinically inspired enhanced explainability and interpretability of an AI tool for BCC diagnosis based on expert annotation." arXiv preprint arXiv:2407.00104, 2024.
- [3] N. Rane, "Transformers for medical image analysis: Applications, challenges, and future scope." Challenges and Future Scope (November 2, 2023), 2023.
- [4] D. Wen, S. M. Khan, A. J. Xu, H. Ibrahim, L. Smith, J. Caballero, L. Zepeda, C. de Blas Perez, A. K.

Table 7

Comparison with Recent State-of-the-Art Studies

Study	Segmentation Method	Classification Method	Dice Score	Accuracy
Akram et al. (2023)	Hybrid CNN	CNN-based classifier	0.92	82.4%
Atiq & Fattah (2025)	Dual-Network Attention Model	Metadata Fusion Network	0.95	84.8%
Abdel-Wahab et al. (2025)	U-Net	EfficientNet	0.96	85.1%
Proposed Framework	U-Net	EfficientNet-B0 + Metadata Fusion	0.9652	86.0%

- Denniston, X. Liu, and R. N. Matin, “Characteristics of publicly available skin cancer image datasets: a systematic review,” *The Lancet Digital Health*, vol. 4, no. 1, pp. e64–e74, 2022.
- [5] S. S. I. Badhon, S. A. Khushbu, N. C. Saha, A. H. Anik, M. A. Ali, and K. T. Hossain, “Explainable AI for Skin Disease Classification Using Grad-CAM and Transfer Learning to Identify Contours,” 2024.
- [6] M. A. Abdou, “Literature review: Efficient deep neural network techniques for medical image analysis,” *Neural Computing and Applications*, vol. 34, no. 8, pp. 5791–5812, 2022.
- [7] T. P. Fowdur and L. Ajageer, “A hybrid online learning-based predictive maintenance system for Industry 4.0,” *Computing*, vol. 107, no. 6, p. 142, 2025.
- [8] G. Dheepak and D. Vaishali, “Brain tumor classification: a novel approach integrating GLCM, LBP, and composite features,” *Frontiers in oncology*, vol. 13, p. 1248452, 2024.
- [9] M. Chowdhury, S. R. Shrima, and M. S. Islam, “Comparative analysis of deep learning architectures for multi-class mineral classification: a study using EfficientNet and ResNet models,” *Earth Science Informatics*, vol. 18, no. 3, p. 485, 2025.
- [10] A. M. Zbrzezny and T. Krzywicki, “Artificial intelligence in dermatology: A review of methods, clinical applications, and perspectives,” *Applied Sciences*, vol. 15, no. 14, p. 7856, 2025.
- [11] S. Chen, Y. Zhang, and Q. Yang, “Multi-task learning in natural language processing: An overview,” *ACM Computing Surveys*, vol. 56, no. 12, pp. 1–32, 2024.
- [12] S. Kumar, A. A. Abdelhamid, and Z. Tarek, “Visualizing the unseen: Exploring GRAD-CAM for interpreting convolutional image classifiers,” *J. Full Length Artic*, vol. 4, pp. 34–42, 2023.
- [13] M. P. Bates, “Visualizing deep learning decisions: grad-CAM-based explainable AI for medical image analysis,” *Journal of Scalable Data Engineering and Intelligent Computing*, pp. 26–33, 2024.
- [14] B. Oliveira, H. R. Torres, P. Morais, F. Veloso, A. L. Baptista, J. C. Fonseca, and J. L. Vilaça, “A multi-task convolutional neural network for classification and segmentation of chronic venous disorders,” *Scientific Reports*, vol. 13, no. 1, p. 761, 2023.
- [15] T. M. Alam, K. Shaukat, W. A. Khan, I. A. Hameed, L. A. Almuqren, M. A. Raza, M. Aslam, and S. Luo, “An efficient deep learning-based skin cancer classifier for an imbalanced dataset,” *Diagnostics*, vol. 12, no. 9, p. 2115, 2022.
- [16] M. E. Atiq and S. A. Fattah, “Towards Explainable Skin Cancer Classification: A Dual-Network Attention Model with Lesion Segmentation and Clinical Metadata Fusion.” arXiv preprint arXiv:2510.17773, 2025.
- [17] A. Akram, J. Rashid, M. A. Jaffar, M. Faheem, and R. U. Amin, “Segmentation and classification of skin lesions using a hybrid deep learning method in the Internet of Medical Things,” *Skin Research and Technology*, vol. 29, no. 11, p. e13524, 2023.
- [18] A. Patil, A. Mehto, and S. Nalband, “Enhancing skin lesion diagnosis with data augmentation techniques: a review of the state-of-the-art,” *Multimedia Tools and Applications*, vol. 84, no. 22, pp. 25325–25364, 2025.
- [19] L. Wang and H. Xiao, “A grid-level segmentation model based on an encoder-decoder structure with multi-source features for crop lodging detection,” *Applied Soft Computing*, vol. 151, p. 111113, 2024.
- [20] X. Polisi, A. Uka, D. Avdiu, and D. Karras, “Efficient and accurate cell image segmentation through optimizer and loss function fine-tuning with quantization and pruning,” *Edelweiss Appl. Sci. Technol*, vol. 9, pp. 1675–1691, 2025.
- [21] M. Alruwaili and M. Mohamed, “An integrated deep learning model with EfficientNet and ResNet for accurate multi-class skin disease classification,” *Diagnostics*, vol. 15, no. 5, p. 551, 2025.
- [22] S. M. Hosseini and M. S. Baghshah, “Dilated balanced cross-entropy loss for medical image segmentation,” *BMC Medical Imaging*, 2026.
- [23] M. W. L. Zheng, “Assessment of Consistency Detection in Mandibular First Premolars Root Canal Morphology Using U-net Model: Evaluation With Dice Coefficient and Intersection Over Union – A Pilot Study,” Master’s thesis, University of Malaya (Malaysia), 2024.
- [24] D. Müller, I. Soto-Rey, and F. Kramer, “Towards a guideline for evaluation metrics in medical image

segmentation,” *BMC research notes*, vol. 15, no. 1, p. 210, 2022.

- [25] A. M. Carrington, D. G. Manuel, P. W. Fieguth, T. Ramsay, V. Osmani, B. Wernly, C. Bennett, S. Hawken, O. Magwood, Y. Sheikh, and M. McInnes, “Deep ROC analysis and AUC as balanced average accuracy, for improved classifier selection, audit, and explanation,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 1, pp. 329–341, 2022.
- [26] T. Hussain and H. Shouno, “Explainable deep learning approach for multi-class brain magnetic resonance imaging tumor classification and localization using gradient-weighted class activation mapping,” *Information*, vol. 14, no. 12, p. 642, 2023.
- [27] Y. Weng, M. Han, H. He, M. Li, L. Yao, X. Chang, and B. Zhuang, “Mask propagation for efficient video semantic segmentation,” *Advances in Neural Information Processing Systems*, vol. 36, pp. 7170–7183, 2023.
- [28] P. K. Parhi, S. Mohanty, P. K. Patra, and A. Mohanty, “Deep Learning Techniques to Detect Brain Tumors Using the EfficientNet-B0 CNN Architecture,” in *Artificial Intelligence in Oncology: Cancer Diagnosis and Treatment, Medical Imaging, and Personalized Medicine*, pp. 115–126, Cham: Springer Nature Switzerland, 2025.
- [29] S. R. Lodhi, S. Arshad, and J. Ali Khan, “A Semi-Supervised Approach utilizing Deep Learning Pseudo-Label Refinement and Confidence Filtering for Enhanced Image Segmentation: An application towards Skin Lesion Analysis,” *MURJET, Pakistan*, pp. 234–243, 2026.