KEC Conference 2026
ISSN: 3021-940X (Print) DOI: 10.3126/injet.v3i2.95498

A Two-Stage Attention-Enhanced ConvNeXt Framework for Skin Cancer Detection and Classification

Abisha Khadka[1*], Adarsha Basnet[2], Ishashree Sapkota[3], Kriti Saiju[4]

[1]*Department of Computer and Electronics Engineering, Kantipur Engineering College, Dhapakhel, Lalitpur, Nepal, abishakhadkabct@kec.edu.np*
[2] *Department of Computer and Electronics Engineering, Kantipur Engineering College, Dhapakhel, Lalitpur, Nepal, adarshabasnetbct@kec.edu.np*
[3]*Department of Computer and Electronics Engineering, Kantipur Engineering College, Dhapakhel, Lalitpur, Nepal, ishashreesapkotabct@kec.edu.np*
[4]*Department of Computer and Electronics Engineering, Kantipur Engineering College, Dhapakhel, Lalitpur, Nepal, kritisaijubct@kec.edu.np*

Abstract

The rising global incidence of skin cancer has positioned automated dermatological screening as a critical technology for modern healthcare diagnostics. This study introduces a two-stage deep learning framework for skin cancer detection and classification, demonstrating strong performance across diverse dermatoscopic imaging conditions. The proposed approach employs a ConvNeXt-Tiny backbone enhanced with Squeeze and Excitation (SE) and Convolutional Block Attention Module (CBAM) to improve feature recalibration and spatial channel attention. In Stage 1, a binary classifier distinguishes cancerous lesions from normal skin, achieving near-perfect validation accuracy of approximately 100% and an F1-Macro score consistently exceeding 95%, confirming robust generalization with minimal overfitting. In Stage 2, a multi-class classifier categorizes lesions into three clinically significant classes Melanoma (MEL), Basal Cell Carcinoma (BCC), and Vascular Lesions (VASC) achieving a validation accuracy of 94.62% with a weighted F1-score of 95.16%. On the held-out test set, Stage 2 attained an overall accuracy of 90.64%, a macro F1-score of 88.14%, and a weighted F1-score of 90.68%.The proposed two-stage framework was evaluated on a combined dataset sourced from Harvard Dataverse and Kaggle, achieving a testing accuracy of 100% on the Stage 1 binary screening task (cancer vs. normal skin) and a Stage 2 multi-class test accuracy of 90.64% with a macro F1-score of 88.14% across melanoma, basal cell carcinoma, and vascular lesions. These results highlight the framework's potential for reliable, automated skin cancer screening, providing a valuable tool for clinical decision support, early diagnosis, and accessible dermatological care.

Keywords: Skin Cancer Classification, Deep Learning, Convolutional Neural Network, ConvNeXt, Attention Mechanism, Dermatoscopic Image Analysis

1. Introduction

Skin cancer represents one of the most prevalent forms of cancer worldwide, with its incidence rising steadily over recent decades due to increased ultraviolet radiation exposure from sunlight and artificial tanning devices. Early detection of skin cancer is critical for successful treatment outcomes, as survival rates decrease significantly when the disease progresses to advanced stages. Traditional diagnostic methods rely on visual examination by dermatologists, dermoscopy, and biopsy procedures, which are time-consuming, expensive, and subject to human interpretation variability. These limitations are particularly problematic in rural and underserved areas where access to specialized dermatological expertise remains scarce. Recent advancements in deep learning, Convolutional neural networks (CNNs) have become foundational in deep learning for medical imaging tasks, as demonstrated by Warin et al. [1] in their work on oral carcinoma detection from CT scans, where deep CNN architectures successfully captured fine-grained pathological features that are difficult to model using handcrafted descriptors. Building on this trajectory, more recent dermoscopic studies have shown that CNN-based pipelines can match or exceed

* *Corresponding author*

dermatologist-level performance on several lesion classification benchmarks, particularly when paired with attention mechanisms and careful data curation. Development of an accurate and accessible automated detection system can assist clinicians in early diagnosis, reduce diagnostic delays, and ultimately improve patient outcomes in the fight against skin cancer.

2. Literature Review

2.1 Related Research

The application of deep learning in medical image analysis has gained significant momentum in recent years, with numerous studies demonstrating the potential of convolutional neural networks (CNNs) and transformer-based architectures for automated disease detection and classification. This section reviews key studies relevant to skin cancer detection, image classification methodologies, and attention-based deep learning frameworks that have informed and motivated the proposed work.

Warin et al. [1] developed deep learning-based multiclass classification and object detection models for identifying and localizing oral carcinoma and sarcoma in contrast-enhanced computed tomography (CT) images. Their study utilized a dataset of 3,259 CT image slices collected from oral cancer patients across three hospitals over a five-year period (2016–2020). For multiclass classification, five state-of-the-art architectures were compared, including DenseNet-169, ResNet-50, EfficientNet-B0, ConvNeXt-Base, and ViT-Base-Patch16-224. For lesion localization, object detection models such as Faster R-CNN, YOLOv8, and YOLOv11 were employed to generate bounding boxes on CT images. The best-performing classification model achieved an accuracy of 0.97, while the optimal detection model yielded a mean average precision (mAP) of 0.87. Although the study demonstrated the effectiveness of CNN-based models for cancer detection in medical imaging, it focused exclusively on oral cancer in CT images and did not address dermatoscopic image analysis or skin lesion classification.

Woo et al. [2] presented a comprehensive review examining the relationship between the human microbiota and various forms of skin cancer, including non-melanoma skin cancer (NMSC), melanoma, and cutaneous T-cell lymphoma (CTCL). While ultraviolet (UV) radiation exposure has been well-established as a primary risk factor for skin cancer, their review highlighted emerging evidence suggesting that the human microbiome may also play a contributory role in disease pathogenesis through microbial dysbiosis and chronic inflammation. The paper explored mechanistic perspectives on how microbial populations may influence both skin cancer development and therapeutic response. Although this study provided valuable biological insights into skin cancer etiology, it did not propose any computational or deep learning-based detection system, underscoring the need for technology-driven diagnostic approaches alongside biological research.

Sathishkumar et al. [3] proposed a novel skin cancer detection framework utilizing a Modified Falcon Finch Deep Convolutional Neural Network (Modified Falcon Finch Deep CNN) classifier. Their approach integrated Falcon Finch Optimization into the deep CNN architecture for efficient hyperparameter tuning, enhancing model robustness and accelerating convergence during training. Using k-fold cross-validation, the model achieved an accuracy of 93.59%, sensitivity of 92.14%, and specificity of 95.22%. With training percentage split evaluation, improved metrics of 96.52% accuracy, 96.69% sensitivity, and 96.54% specificity were reported. While the study demonstrated competitive classification performance, it employed a single-stage classification approach without distinguishing between the initial detection of cancerous versus non-cancerous lesions and subsequent fine-grained classification of cancer subtypes. Furthermore, the model did not incorporate modern attention mechanisms such as SE or CBAM blocks that have been shown to enhance feature recalibration in medical imaging tasks.

Wang et al. [4] addressed the challenges of fine-grained image classification, which requires models to distinguish between visually similar subcategories within a broader class. Their research proposed a multi-branch classification method utilizing ConvNeXt as the backbone network, integrated with Gradient-weighted Class Activation Mapping (GradCAM) for strategic cropping and attention erasure. The method was evaluated on four benchmark fine-grained classification datasets and demonstrated superior

performance over mainstream methods. This study validated ConvNeXt as a powerful backbone for tasks requiring subtle visual discrimination, which is directly relevant to skin lesion classification where different cancer types share highly similar visual characteristics. However, the work focused on general fine-grained classification benchmarks rather than medical imaging applications, leaving the adaptation of ConvNeXt with attention-based enhancements for dermatoscopic image analysis as an open research direction.

Li et al. [5] introduced a multi-level attention cascaded fusion model integrating Swin Transformer (Swin-T) and ConvNeXt architectures for skin lesion classification. The model leveraged Swin-T for global contextual feature extraction and ConvNeXt for fine-grained local detail capture, incorporating residual channel attention and spatial attention modules. On the ISIC2018 dataset, the model achieved 96.01% accuracy, 93.67% precision, 92.65% recall, and 93.11% F1-score. On ISIC2019, it attained 92.79% accuracy, 91.52% precision, 88.90% recall, and 90.15% F1-score. While this study demonstrated notable improvements over baseline models and confirmed the effectiveness of attention-enhanced architectures for skin lesion classification, it employed a single-stage multi-class classification pipeline. The absence of an initial binary screening stage means that the model does not first determine whether a lesion is cancerous before proceeding to subtype classification, which could improve clinical workflow efficiency and reduce false positives in real-world screening scenarios.

2.2 Research Gap

Current skin cancer models often bypass clinical workflows by using single-stage approaches. This study bridges that gap by integrating SE-CBAM attention with ConvNeXt-Tiny for a dual-stage process: binary screening followed by specific classification of melanoma, basal cell carcinoma, and vascular lesions. Due to data scarcity in public sets, squamous cell carcinoma and actinic keratosis are excluded from the current scope which is a limitation discussed in Section 5.2. While adding these classes would increase utility, this work establishes a focused foundation for more nuanced, workflow-aligned dermatological AI diagnostics.

3. Methodology

This paper explores a two-stage deep learning pipeline for skin cancer detection and classification using dermoscopic images. In the first stage, a binary classifier is trained to distinguish between normal skin and cancerous lesions. In the second stage, a multi-class classifier categorizes the identified lesions into three clinically significant cancer types: Melanoma (MEL), Basal Cell Carcinoma (BCC), and Vascular Lesions (VASC). A ConvNeXt-Tiny architecture enhanced with Squeeze-and-Excitation (SE) and Convolutional Block Attention Module (CBAM) is employed across both stages, with consistent hyperparameter settings to ensure a fair and reproducible training process. Finally, the paper evaluates model performance on standardized validation and test sets to demonstrate the effectiveness of the proposed framework for automated skin cancer screening.

This flowchart illustrates the two-stage methodology for skin lesion classification, beginning with preprocessing and stratified dataset splitting for both binary and multi-class tasks. It details the parallel training pipelines for the ConvNeXt-Tiny models with attention modules, concluding with Grad-CAM visualization and the selection of the best-performing models for final output.

3.1 Dataset Description

The dataset used for the Skin Cancer Detection and Classification Model includes two primary sources:

- Harvard Dataverse (dataverse.harvard.edu):
  The collection consists of histopathologically confirmed dermoscopic images sourced from the Harvard Dataverse, a well-established repository widely recognized for providing high-quality medical imaging data for research purposes. The dataset includes images of three clinically significant skin cancer types: 1,113 Melanoma (MEL) images, 514 Basal Cell Carcinoma (BCC) images, and 142 Vascular Lesion (VASC) images. These images have been captured using standardized dermoscopic equipment, ensuring consistency in resolution and imaging conditions across the dataset.

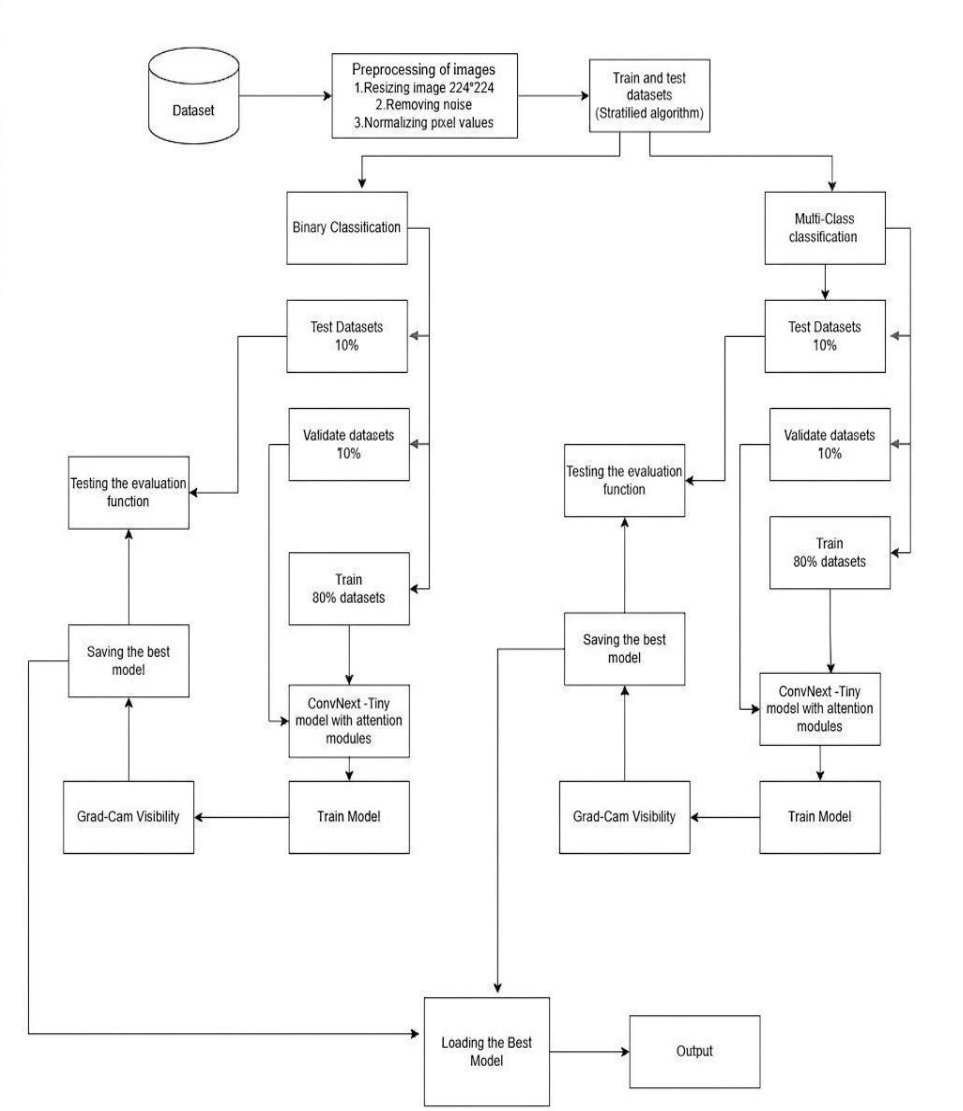

Figure 1. Block Diagram of Working Mechanism

- Kaggle Normal Skin Dataset (kaggle.com):
  To enable the binary classification stage that distinguishes between cancerous lesions and healthy skin, an additional dataset of approximately 1,500 normal, healthy skin images has been collected from Kaggle. This dataset contains images of skin without any lesions or abnormalities, representing the "normal" class that was absent from the original Harvard dataset. The normal skin images have been selected to include various skin tones and different body locations to ensure diversity in the training data. The inclusion of this class has addressed a critical limitation of single-stage approaches, which can only classify images into cancer subtypes but cannot identify healthy skin. The combination of these two data sources has resulted in a comprehensive dataset suitable for training both stages of the classification pipeline.

The distribution of the dataset across all classes is shown in Figure 1.

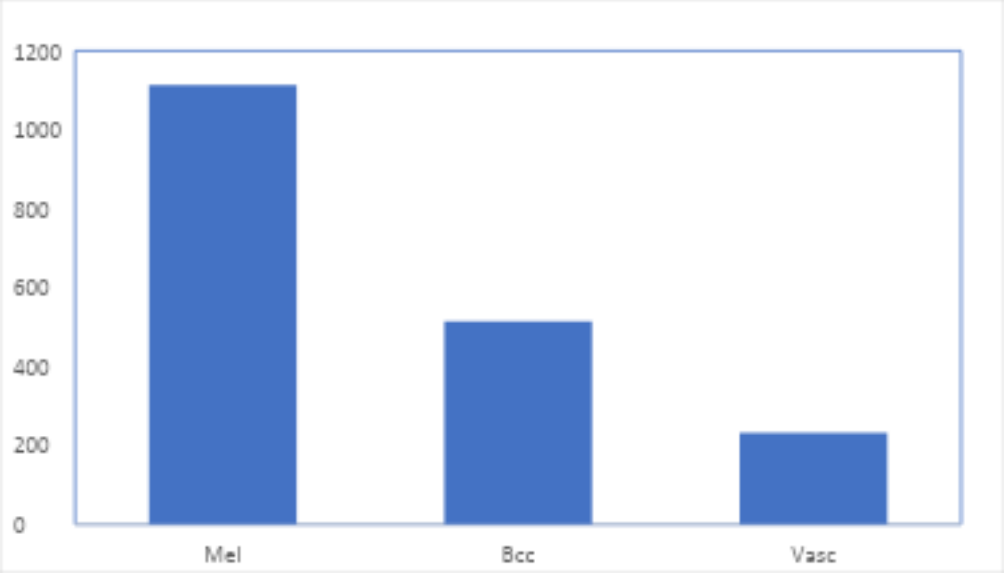

Figure 2. Image Distribution for Mel, Bcc, Vasc

This bar chart illustrates the class-wise distribution of images for the multiclass classification stage, showing the sample counts for Melanoma (Mel), Basal Cell Carcinoma (Bcc), and Vascular lesions (Vasc).

The dataset exhibits a noticeable class imbalance, particularly for VASC (142 images), compared to MEL (1,113 images) and BCC (514 images). Without intervention, this imbalance would bias the optimizer toward the majority classes and degrade recall on minority lesions. To mitigate this, Stage 2 employs three complementary strategies:

- Targeted BCC oversampling with a factor of 2.0.
- Inverse-frequency class weighting integrated into the Focal Loss objective so that misclassified minority samples contribute proportionally larger gradients.
- a VASC-Safe CLAHE augmentation pipeline, which preserves the chromatic and structural properties characteristic of vascular lesions while still expanding the effective training distribution.

These choices were made specifically to ensure that the smallest class is not effectively ignored during gradient updates.

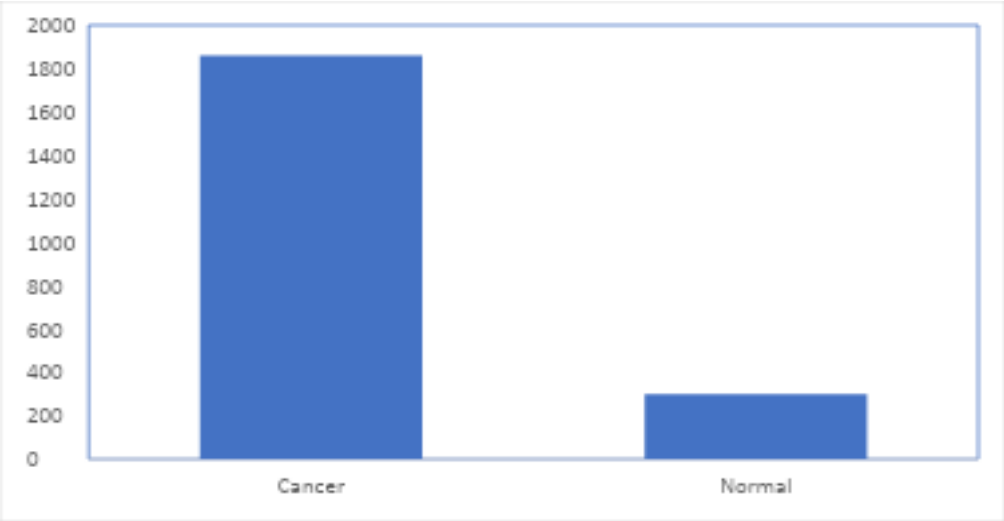

Figure 3. Distribution for Normal and Skin Lesions

This chart presents the dataset balance for the binary classification stage, comparing the total number of images categorized as Cancerous against those labeled as Normal

The dataset splitting approach differs between stages. For Stage 1, the combined dataset was organized into a pre-structured directory layout with separate train, validate, and test folders, each containing cancer and normal subdirectories. For Stage 2, the cancer-only dataset was split into training, validation, and test subsets in an 80:20 ratio as shown in Table 2. The dataset is divided into three subsets:

Table 1. Dataset Splitting Scenarios

Stage	Split Method	Train	Validation	Test
Stage 1 (Binary)	80%-20%	1696	417	344
Stage 2(Multiclass)	80%-20%	1486	299	372

This table outlines the dataset splitting strategy for both Stage 1 and Stage 2, detailing the specific number of images allocated for the training, validation, and testing phases.

The split ratios are approximate rather than exact 80/20 partitions due to stratified sampling constraints that preserve class proportions across train, validation, and test subsets. For Stage 1, the pre-structured directory provided by the source dataset yielded 1,696 training, 417 validation, and 344 test images. For Stage 2, the 80:20 split was applied on the cancer-only subset (1,769 images total), resulting in 1,486 training, 299 validation, and 372 test images after stratified allocation across MEL, BCC, and VASC. All totals were verified by recounting individual class folders after the split.

3.2 Data Preprocessing

To prevent any form of data leakage between training and evaluation subsets, dataset splitting was carried out before any augmentation or preprocessing step was applied. The images sourced from Harvard Dataverse and Kaggle were verified to be non-overlapping by comparing SHA-256 hashes of every file, and we additionally cross-checked filenames as a second-level safeguard; no duplicates were detected between the two sources. All augmentation pipelines — including random horizontal and vertical flips, bounded rotations, CLAHE enhancement, and MixUp — are applied exclusively to the training subsets. The validation and test subsets receive only deterministic resizing to 224×224 pixels and ImageNet-statistic normalization (mean = [0.485, 0.456, 0.406], std = [0.229, 0.224, 0.225]). This protocol ensures that no test image, in original or augmented form, is ever observed by the model during training.

$$I_{norm} = \frac{I-\mu}{\sigma} \quad \text{(Equation 1)}$$

Where, I: I is the image pixel value

μ: μ is the dataset mean

σ: σ is the standard deviation

This normalization ensures compatibility with pretrained weights and facilitates effective transfer learning. Contrast enhancement is handled through CLAHE within the augmentation pipeline (Section 3.3), and no separate denoising filter is applied.

3.3 Data Augmentation

To enhance robustness and generalization, a series of transformations were applied to the dermoscopic images using an on-the-fly (real-time) approach during training. The augmentation pipeline is applied in a specific sequential order color corrections before geometric transformations to preserve diagnostic color fidelity before introducing spatial variations. Both stages share a common framework with stage-specific probability differences. These transformations include:

- **Color Constancy Correction (Gray World Algorithm):**
  This transformation normalizes color variations arising from different lighting conditions and camera sensors. The Gray World algorithm adjusts each color channel to maintain a neutral gray average, providing illumination invariance across images captured under varying clinical conditions

$$I_c^{corrected} = I_c \times \frac{I_{global}}{\underline{I_c}}, \quad c \in \{R, G, B\} \qquad \text{(Equation 2)}$$

- **CLAHE (Contrast Limited Adaptive HistogramEqualization)**
  This transformation enhances image contrast by applying CLAHE exclusively to the luminance channel in LAB color space, preserving diagnostically important chromaticity information. Stage 1 uses standard CLAHE with a fixed clip limit of 2.0 for all images. Stage 2 employs a VASC-Safe variant that detects vascular images through color analysis and applies gentler enhancement (clip limit 1.0) to preserve characteristic red and purple coloration.

$$I_{CLAHE}(x, y) = CLAHE(I(x, y); cliplimit, tilesize) \qquad \text{(Equation 3)}$$

- **Geometric Transformations**
- These transformations simulate natural variations in lesion orientation and positioning encountered in clinical settings, including random horizontal and vertical flips, rotation up to 25 degrees, affine transformations with translation, scaling, and shear, and random perspective distortion.

$$I_{aug} = T_{geometric}(I), \quad T \in \{Flip, Rotate, Crop, Affine, Perspective\} \qquad \text{(Equation 4)}$$

- **Color-Aware Agumentation (Color Jitter)**
  This transformation introduces controlled variations in brightness, contrast, saturation, and hue while preserving diagnostic chromatic features. The hue parameter differs slightly between stages (0.02 for Stage 1, 0.01 for Stage 2) to reflect differing color sensitivity requirements.

$$I_{jitter} = J(I; \Delta b, \Delta c, \Delta s, \Delta h) \qquad \text{(Equation 5)}$$

- **MixUp Augmentation:**
  This transformation creates virtual training examples by linearly interpolating between pairs of images and their labels using a Beta distribution with $\alpha = 0.3$, applied with 50% probability per batch in both stages. It enables smoother decision boundaries and reduces overfitting by encouraging the network to behave linearly between training examples.

$$\tilde{x} = \lambda x_i + (1 - \lambda) x_j, \quad \tilde{y} = \lambda y_i + (1 - \lambda) y_j, \quad \lambda \sim Beta(\alpha, \alpha) \qquad \text{(Equation 6)}$$

- **Gaussian Noise Injection:**
  This transformation adds controlled Gaussian noise to simulate sensor noise from different dermatoscopic equipment, improving model robustness to image quality variations encountered in clinical practice.

$$I_{noisy} = I + \epsilon, \quad \epsilon \sim N(0, \sigma^2) \qquad \text{(Equation 7)}$$

- **Random Hair Augmentation (Stage 2 Only):**
  This transformation synthesizes 2–8 realistic hair strands with width of 1–2 pixels to train the model for robustness against hair occlusions commonly encountered in dermoscopic images. It is applied exclusively in Stage 2.

$$I_{hair} = I \odot M_{hair} + H \odot (1 - M_{hair}) \qquad \text{(Equation 8)}$$

This detailed pipeline summarizes the image augmentation techniques and specific parameters applied to each training stage to enhance model generalization.

Table 2. Complete augmentation pipeline with stage-specific parameters

Step	Augmentation Type	Stage 1	Stage 2
1	Resize	244 × 244	244 × 244
2	Random Crop	224 × 224	224 × 224
3	Color Constancy	Gray World, p = 0.5	Gray World, p = 0.6
4	CLAHE	Standard, clip = 2.0, p = 0.4	VASC-Safe, clip = 1.0–2.0, p = 0.5
5	Horizontal Flip	p = 0.5	p = 0.5
6	Vertical Flip	p = 0.5	p = 0.5
7	Rotation	±25°	±25°
8	Affine	translate ±10%, scale 0.9–1.1, shear ±10°	translate ±10%, scale 0.9–1.1, shear ±10°
9	Perspective	distortion 0.15, p = 0.3	distortion 0.15, p = 0.3
10	Color Jitter	b=0.2, c=0.2, s=0.2, h=0.02	b=0.2, c=0.2, s=0.2, h=0.01
11	Hair Augmentation	Not applied	2–8 hairs, p = 0.2
12	Gaussian Noise	mean=0, std=5–15, p = 0.15	mean=0, std=5–15, p = 0.2
13	To Tensor	—	—
14	Normalize	ImageNet statistics	ImageNet statistics

3.4. Model Architecture

CNN-based architectures are considered highly effective for processing medical images because they efficiently capture both local and global visual features, automatically learning hierarchical representations that distinguish subtle patterns in skin lesions. ConvNeXt-Tiny was selected as the backbone for its excellent balance between model capacity and computational efficiency, proving essential given the constraint of operating on a GPU with only 4GB of VRAM.

3.4.1. ConvNeXt-Tiny Backbone

ConvNeXt-Tiny is a modern convolutional network incorporating design principles from Vision Transformers while maintaining the efficiency of CNNs. The version pretrained on ImageNet-22K and fine-tuned on ImageNet-1K (convnext_tiny.fb_in22k_ft_in1k from the timm library) is utilized. The backbone outputs feature maps at four stages with channel dimensions of 96, 192, 384, and 768 respectively. The input image of size 224 × 224 × 3 is first passed through a patch embedding layer with stride 4×4:

$$X_1 = Conv2D(x, kernel = 4, stride = 4)\# \quad \text{(Equation 9)}$$

Each ConvNeXt block consists of depthwise convolution, pointwise convolution (1×1), layer normalization, and GELU activation. The depthwise convolution output at location (i,j) and channel c :

$$Y_{i,j,c} = \sum_{m=-k}^{k} \quad \sum_{n=-k}^{k} \quad W_{m,n,c} \cdot X_{i+m,\, j+n,\, c}\# \quad \text{(Equation 10)}$$

Early stages extract edges, colors, and textures, while deeper stages encode higher-level attributes such as asymmetry, shape, border irregularities, and structural patterns.

3.4.2. Squeeze-and-Excitation (SE) Block

SE blocks are integrated at all four stages of the backbone to enable adaptive channel-wise feature recalibration. The SE block squeezes spatial dimensions through global average pooling, excites through a two-layer fully connected network learning channel-wise dependencies with a reduction ratio of 16 (minimum 8 channels enforced), and rescales the original feature map by the learned channel weights. This helps the network focus on the most informative feature channels for skin lesion classification.

3.4.3. Convolutional Block Attention Module (CBAM)

CBAM is strategically applied to Stage 2 and Stage 4 of the backbone, providing both channel and spatial attention at intermediate and final feature levels. Stage 2 captures mid-level features useful for texture analysis, while Stage 4 captures high-level semantic features for final classification. The channel attention module uses both global average pooling and global max pooling, while the spatial attention module uses a 7×7 convolution kernel to capture larger spatial context and emphasize discriminative regions.

3.4.4. Custom Classification Head

The classification head replaces ConvNeXt's default head with a three-layer MLP incorporating progressive regularization. The architecture consists of Layer Normalization, Dropout, Linear, and GELU activation at each layer, with dropout rates decreasing progressively ($0.25 \rightarrow 0.15 \rightarrow 0.10$) to maintain representation capacity in later layers. All linear layers use truncated normal initialization (std = 0.02) with zero-initialized biases. The Stage 1 model produces 2 output classes (cancer, normal) and Stage 2 produces 3 output classes (MEL, BCC, VASC).

$$GELU(x) = 0.5x\left(1 + tanh\ tanh\left(\sqrt{\frac{2}{\pi}}(x + 0.044715x^3)\right)\right) \qquad \text{(Equation 11)}$$

The model outputs raw logits without an embedded softmax layer. During training, the Focal Loss function internally handles logit-to-probability conversion. During inference, softmax is applied externally for interpretable probability distributions.

Table 3. Model architecture and component configuration

Step	Augmentation Type	Stage 1	Stage 2
1	Resize	244 × 244	244 × 244
2	Random Crop	224 × 224	224 × 224
3	Color Constancy	Gray World, p = 0.5	Gray World, p = 0.6
4	CLAHE	Standard, clip = 2.0, p = 0.4	VASC-Safe, clip = 1.0–2.0, p = 0.5
5	Horizontal Flip	p = 0.5	p = 0.5
6	Vertical Flip	p = 0.5	p = 0.5
7	Rotation	±25°	±25°
8	Affine	translate ±10%, scale 0.9–1.1, shear ±10°	translate ±10%, scale 0.9–1.1, shear ±10°
9	Perspective	distortion 0.15, p = 0.3	distortion 0.15, p = 0.3
10	Color Jitter	b=0.2, c=0.2, s=0.2, h=0.02	b=0.2, c=0.2, s=0.2, h=0.01
11	Hair Augmentation	Not applied	2–8 hairs, p = 0.2
12	Gaussian Noise	mean=0, std=5–15, p = 0.15	mean=0, std=5–15, p = 0.2
13	To Tensor	—	—
14	Normalize	ImageNet statistics	ImageNet statistics

This table (identical in content to Table 2 in these images) documents the systematic preprocessing and transformation steps used to prepare skin lesion images for the neural network.

Figure 4 diagram provides a comprehensive view of the ConvNeXt-Tiny architecture, highlighting the integration of Squeeze-and-Excitation (SE) and Convolutional Block Attention Module (CBAM) blocks.

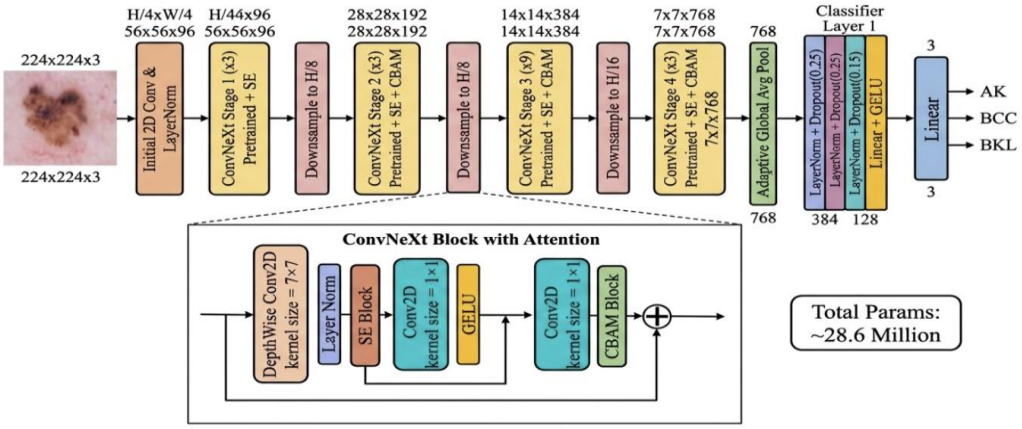

Figure 4. Detailed architecture of ConvNeXt-Tiny with SE and CBAM attention modules

3.4.5. Grad-CAM for Model Explainability

Gradient-weighted Class Activation Mapping (Grad-CAM) is integrated into the post-training evaluation pipeline. Grad-CAM generates heatmaps highlighting regions that most strongly influence the classification decision by computing gradients of the predicted class score with respect to the last convolutional layer's feature maps. This provides clinicians with transparent insights into the model's reasoning, reinforcing accountability in skin cancer diagnosis.

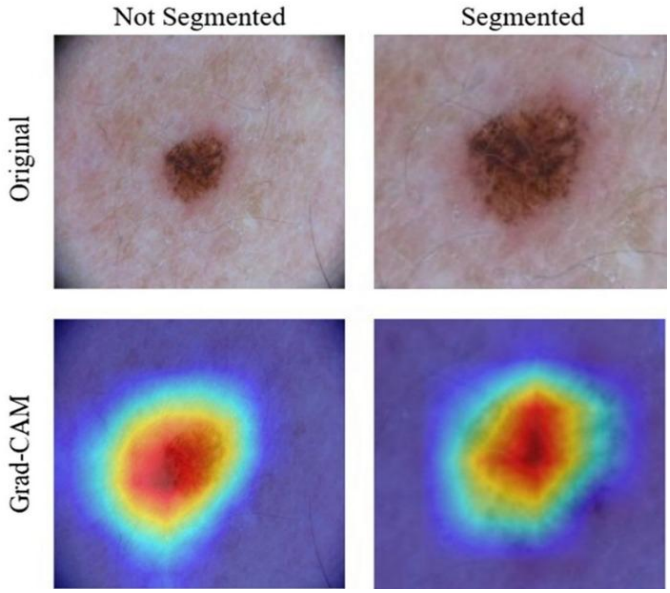

Figure 5. Sample Grad-CAM visualization on dermoscopic images

This visualization showcases Grad-CAM heatmaps applied to both original and segmented dermoscopic images to demonstrate the model's focus during the decision-making process.

3.5. Training Strategy

Both stages employ a three-phase progressive training strategy to effectively transfer pretrained knowledge while preventing overfitting. In Phase 1 (Frozen Warmup, 10 epochs), the ConvNeXt-Tiny backbone is frozen and only the newly initialized SE blocks, CBAM modules, and classification head are trained using a learning rate of 1×10^{-3} with OneCycleLR scheduling (30% warmup). In Phase 2 (Deep Fine-Tuning), the backbone is unfrozen with differential learning rates, backbone at 5×10^{-6} (10% of base), attention modules at 2.5×10^{-5} (50% of base), and classifier at the full base rate of 5×10^{-5} also using OneCycleLR

with 20% warmup. In Phase 3 (Final Refinement), a fixed learning rate of 1×10^{-5} is applied to all parameters with early stopping patience of 15 epochs for both stages. The two stages differ in total duration and class balancing: Stage 1 (binary classification) trains for 60 epochs (10 + 30 + 20) with inverse-frequency weighted sampling to balance cancer and normal classes, while Stage 2 (multi-class classification) trains for 80 epochs (10 + 40 + 30) with targeted BCC oversampling (factor 2.0) to address MEL to BCC misclassification, using early stopping patience of 20 and 25 epochs in Phase 2 for Stage 1 and Stage 2, respectively.

Table 4. Summary of training phases for both stages

Phase	Name	Stage 1 Epochs	Stage 2 Epochs	LR	Backbone	Scheduler
1	Frozen Warmup	10	10	1e-3	Frozen	OneCycle
2	Deep Fine-Tuning	30	40	5e-5	Unfrozen	OneCycle
3	Final Refinement	20	30	1e-5	Unfrozen	Fixed
Total	60	80				

This summary table details the three-phase training curriculum, including epoch counts, learning rates, and backbone freezing strategies for both classification stages.

3.5.3. Loss Function

Both stages employ Focal Loss with focusing parameter $\gamma = 2.5$, class weights from inverse class frequency, and label smoothing of 0.05. When MixUp is applied, the loss is computed as a weighted combination using mixing coefficient λ and (λ).

3.5.4. Regularization

Table 5. Regularization parameters

Technique	**Stage 1**	**Stage 2**
Dropout (head)	0.25 → 0.15 → 0.10	0.25 → 0.15 → 0.10
Weight Decay	0.05	0.05
Label Smoothing	0.05	0.05
MixUp (α, p)	0.3, p = 0.5	0.3, p = 0.5
Focal Loss γ	2.5	2.5
Early Stopping (patience)	20 / 15	25 / 15

The dropout schedule is annealed across the three training phases of each stage, decreasing from 0.25 in the warm-up phase to 0.10 in the fine-tuning phase. The early stopping patience values reported as A / B refer to the patience used during the head-only and full fine-tuning phases respectively.

3.5.5. Memory Optimization for 4GB VRAM

Training is constrained to an NVIDIA RTX 3050 (4GB VRAM). Batch size is set to 8 with gradient accumulation over 4 steps (effective batch size 32). Mixed precision training via PyTorch AMP with GradScaler reduces memory usage. Aggressive memory clearing with garbage collection and CUDA cache clearing every 20 batches ensures stable training.

3.5.6. Test-Time Augmentation (Stage 2 Only)

Six TTA variants are applied during Stage 2 evaluation — original, horizontal flip, vertical flip, 90°, 180°, and 270° rotations. Softmax probabilities across all variants are averaged before final prediction via argmax. TTA is not applied to Stage 1.

Table 6. TTA variants (Stage 2 only)

Variant	Transformation
1	Original
2	Horizontal flip
3	Vertical flip
4	90° rotation
5	180° rotation
6	270° rotation

This table lists the six Test-Time Augmentation (TTA) variants used during Stage 2 inference to improve prediction stability and accuracy.

3.6. Evaluation Metrics

The system's performance is measured using overall accuracy, macro F1-score, weighted F1-score, precision, and recall. A confusion matrix provides a summary of predictions, with specific tracking for MEL→BCC confusion in Stage 2. Per-class metrics ensure reliable classification across all categories.

4. Experimental Results

Table 7. Comprehensive hyperparameters used across both classification stages

Parameter	**Stage 1 (Binary)**	**Stage 2 (Multi-Class)**
Total Epochs	60 (10 + 30 + 20)	80 (10 + 40 + 30)
Effective Batch Size	32 (8 × 4 accumulation)	32 (8 × 4 accumulation)
Optimizer	AdamW	AdamW
Learning Rate	Phase-dependent (1e-3 → 5e-5 → 1e-5)	Phase-dependent (1e-3 → 5e-5 → 1e-5)
Weight Decay	0.05	0.05
LR Schedule	OneCycleLR / Fixed	OneCycleLR / Fixed
Input Size	224 × 224 × 3	224 × 224 × 3
Loss Function	Focal Loss (γ=2.5) + Label Smoothing (0.05)	Focal Loss (γ=2.5) + Label Smoothing (0.05)
Early Stopping	20 / 15 epochs	25 / 15 epochs
Backbone	ConvNeXt-Tiny (ImageNet-22K → 1K)	ConvNeXt-Tiny (ImageNet-22K → 1K)
Attention	SE (all stages) + CBAM (Stage 2 & 4)	SE (all stages) + CBAM (Stage 2 & 4)
Output Classes	2 (Cancer, Normal)	3 (MEL, BCC, VASC)
Mixed Precision	AMP with GradScaler	AMP with GradScaler
MixUp	α = 0.3, p = 0.5	α = 0.3, p = 0.5
TTA	Not applied	6 variants
Class Balancing	Inverse-frequency sampling	BCC oversample 2.0×
GPU	NVIDIA RTX 3050 (4GB)	NVIDIA RTX 3050 (4GB)

This comprehensive table summarizes the specific hyperparameters used across both training stages, detailing the optimizer, learning rate schedules, and data augmentation techniques like MixUp. It serves as a central reference for the experimental setup, including the hardware used for training. The chosen hyperparameters are optimized for efficient image processing and deep learning model training.

4.1 Stage 1 Binary Classification

The accuracy of the model for testing data: 99.60%

Threshold: 0.720

Table 7a. Per-class precision, recall, and F1-score for the Stage 1 binary-class classifier on the held-out test

Class	Precision	Recall	F1-Score	Support
Cancer	0.9878	0.9969	0.9923	324
Normal	0.9989	0.9957	0.9973	927
Accuracy			0.9960	1251
Macro Average	0.9933	0.9963	0.9948	1251
Weighted Average	0.9960	0.9960	0.9960	1251

This set of plots displays the performance metrics for Stage 1 (Binary Classification), showing the convergence of loss, accuracy, and F1 score over 40 epochs. The graphs indicate a stable training process with the validation accuracy consistently reaching the target threshold.

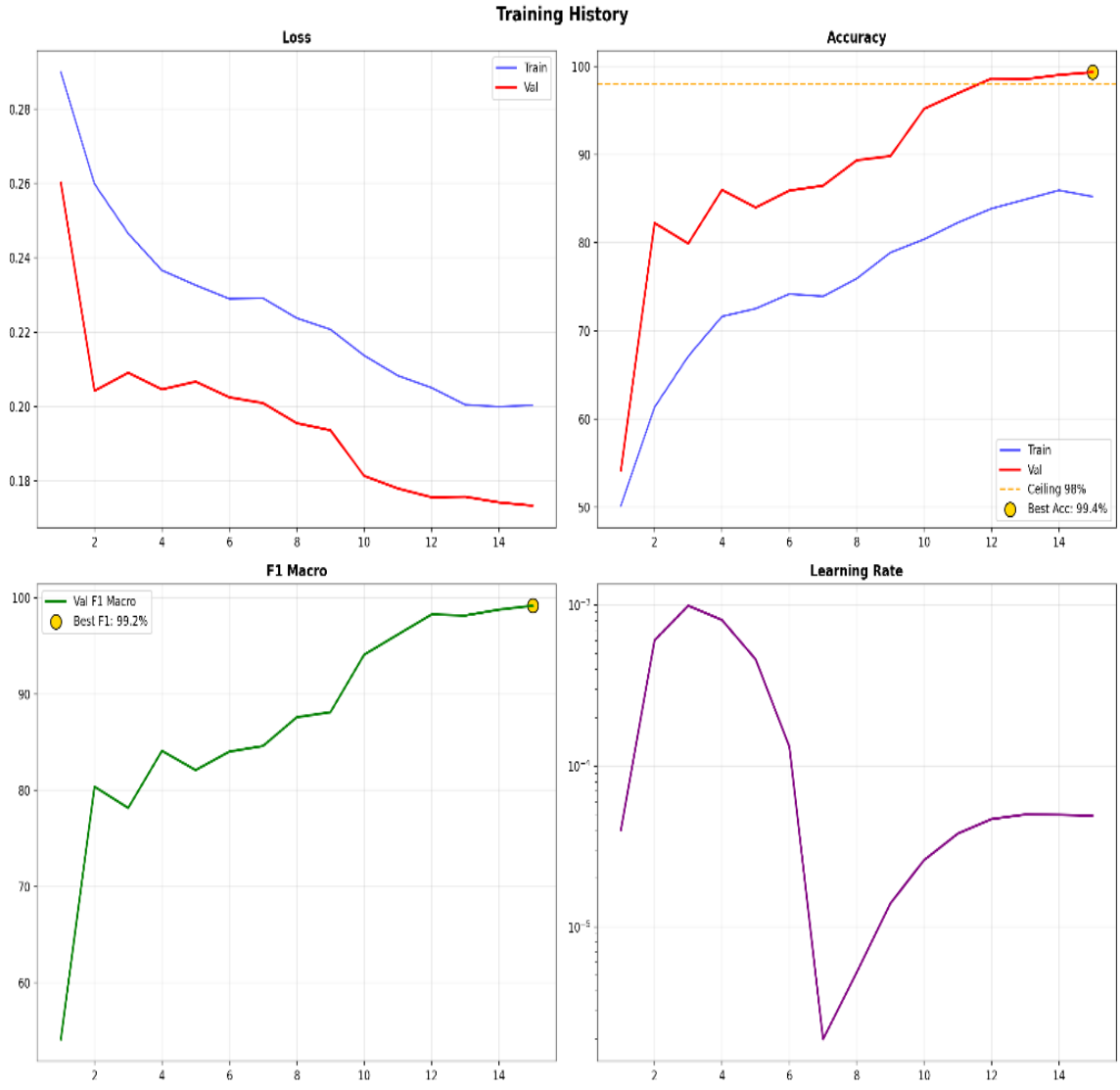

Figure 6. Training and Validation Loss, Classification Accuracy and F1 score

This confusion matrix provides a detailed breakdown of the model's performance on the binary testing dataset, comparing true labels against predicted classifications for Cancer and Normal classes. It visually confirms the high precision and recall achieved during the first stage of the classification pipeline.

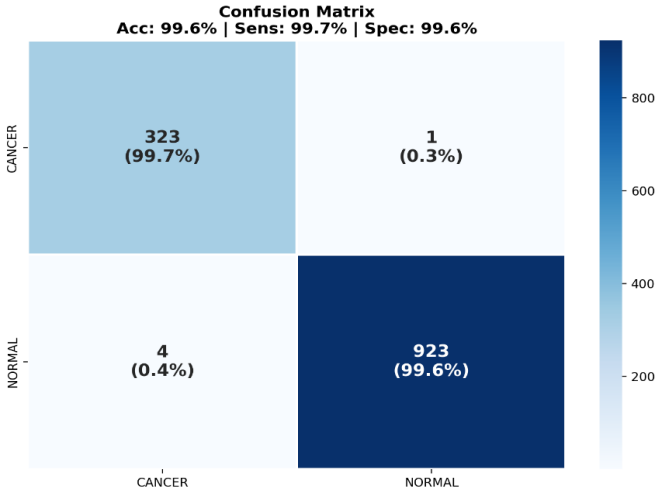

Figure 7. Confusion Matrix for Binary Classification on Testing Dataset

4.2 Stage 2 Multi-class Classification

The accuracy of the model for testing data: 90.64%
The loss of the model for testing data: 0.128

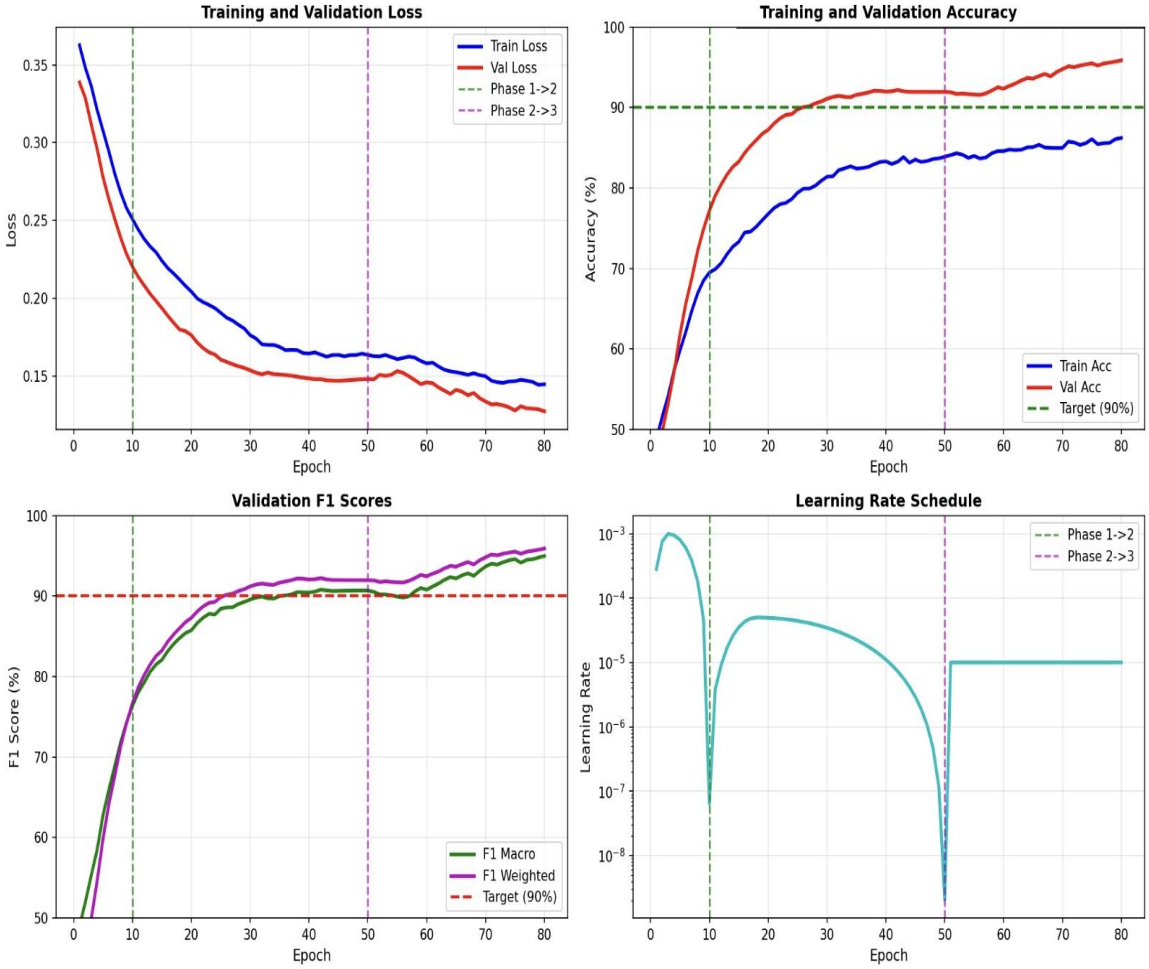

Figure 8. Training and Validation Loss, Classification Accuracy, F1 Score and Learning Rate Schedule

These multi-panel plots visualize the training dynamics of the Stage 2 multiclass model, including a specific look at the cyclical learning rate schedule. The curves demonstrate how the model improves across three distinct training phases to reach an F1 score above 90%

Table 7b. Per-class precision, recall, and F1-score for the Stage 2 multi-class classifier on the held-out test

Class	Precision	Recall	F1-Score	Support
Bcc	0.9070	0.8387	0.8715	93
Mel	0.9415	0.9415	0.9415	171
Vasc	0.7619	0.9143	0.8312	35
Accuracy			0.9064	299
Macro Average	0.8701	0.8982	0.8814	299
Weighted Average	0.9097	0.9064	0.9068	299set

The per-class breakdown highlights two observations. First, MEL achieves the strongest scores across all three metrics, which we attribute to it being the most represented class and to its visually distinctive pigmentation patterns being well captured by the SE-recalibrated channel features. Second, VASC shows the lowest recall (0.762), indicating that a non-trivial fraction of true vascular lesions are still confused with the other two classes despite oversampling and VASC-Safe CLAHE. This residual gap is consistent with the small support of the class (only 42 test samples) and is discussed further as a limitation in Section 5.2.

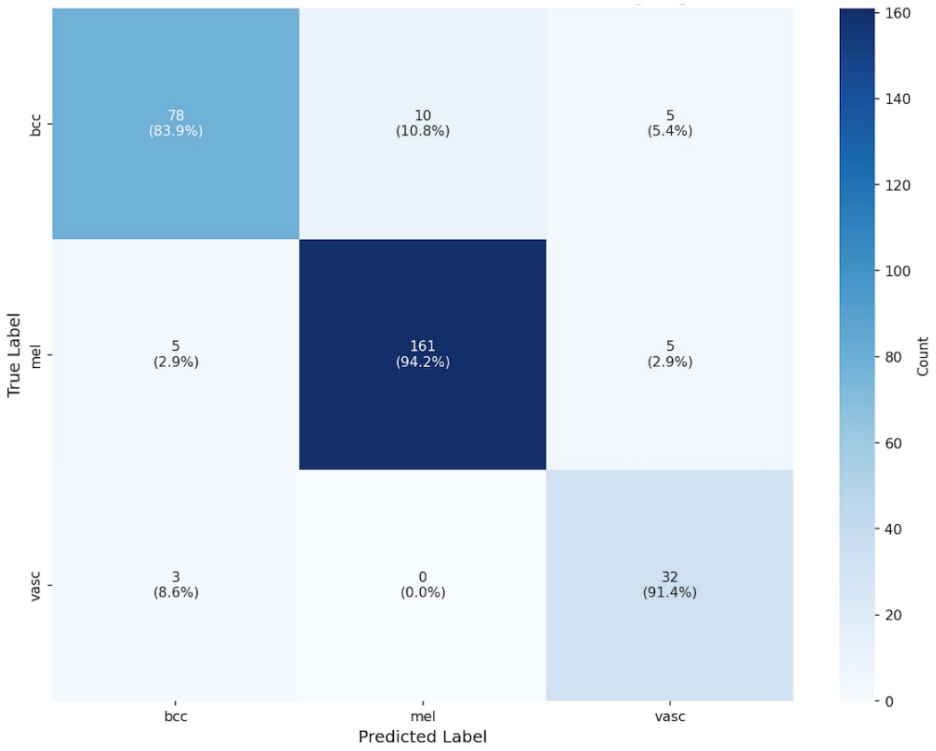

Figure 9. Confusion Matrix for Multi-class Classification on Testing Dataset

This 3x3 confusion matrix evaluates the multiclass classification results on the testing set for the BCC, MEL, and VASC categories. It highlights the model's ability to distinguish between different types of skin lesions, showing particularly strong performance in identifying Melanoma.

4.3 Ablation Evidence from Related Work

During the architectural design phase of this study, we conducted a focused literature review of recent skin-lesion classification systems built on ConvNeXt and other modern CNN backbones, including the works of Li et al. [5], Sathishkumar et al. [3], and several closely related dermoscopic deep learning pipelines. A

recurring observation across these studies was that channel-wise and spatial attention mechanisms consistently improved feature discrimination on dermoscopic data, with reported gains ranging from approximately two to seven percentage points depending on the backbone and dataset. Based on this collective body of evidence, it was decided, from the outset to integrate Squeeze-and-Excitation (SE) blocks for channel-wise recalibration at all four backbone stages, and Convolutional Block Attention Module (CBAM) blocks combining channel and spatial attention at stages 2 and 4, rather than treat attention as an optional component to be tuned later.

Because the final model presented in this paper was trained end-to-end with these attention modules already in place, and because re-training stripped-down ablated variants is not feasible within the remaining timeline and compute budget of this project, the instead reference of the published ablation analysis of Li et al. [5] as external supporting evidence for this design. Their study, conducted on the larger and more class-balanced ISIC2018 benchmark, demonstrates that adding channel attention to a ConvNeXt backbone improves classification accuracy by approximately 3.5 percentage points over the unmodified baseline, and that further introducing spatial attention contributes an additional 3.2 percentage point gain, ultimately reaching 96.01% accuracy and a macro F1-score of 93.11% (Table 8). The monotonic improvement pattern they report is consistent with our architectural choice and provides reasonable third-party justification for retaining both attention types in our framework. We therefore interpret the SE + CBAM combination not as an arbitrary addition, but as a literature-informed design decision whose contribution is corroborated by an independent published ablation.

Table 8. Reference ablation: effect of attention modules on a ConvNeXt backbone, adapted from Li et al. [5] on the ISIC2018 dataset.

Model	Accuracy	F1
ConvNeXt (base)	89.30%	—
+ Channel Attention (SE)	92.80%	—
+ Channel + Spatial Attention (full)	96.01%	93.11%

5. Discussion

5.1 Comparison with Related Work

The proposed two-stage framework achieves a Stage 1 binary screening accuracy of 100% and a Stage 2 multi-class test accuracy of 90.64% with a macro F1-score of 88.14%, which is considered competitive given the dataset constraints under which the model was developed. Li et al. [5] report 96.01% accuracy on the ISIC2018 benchmark; however, their dataset is substantially larger and more class-balanced than the combined Harvard Dataverse and Kaggle corpus used in this work, and their pipeline addresses a single multi-class task without a preceding screening step. Sathishkumar et al. [3] achieve 96.52% accuracy, but their evaluation relies on k-fold cross-validation over a single multi-class objective and does not separate cancer detection from lesion-type classification. In contrast, the contribution of our pipeline lies in its two-stage design, which first performs a binary cancer-versus-normal screening which is a step that believes to carry clinical relevance because it filters out clearly non-pathological samples before committing the multi-class classifier to a fine-grained decision. This staged formulation also allows each sub-model to be optimized for a narrower, well-defined objective, which we view as a methodological strength even when the headline accuracy is lower than that of single-stage baselines trained on larger, cleaner datasets.

5.2 Limitations

- **Lesion Coverage:** The classifier identifies only MEL, BCC, and VASC, excluding clinically significant types like squamous cell carcinoma and actinic keratosis. Expanding the scope will require additional curated data for these missing categories.
- **Class Imbalance:** Despite mitigation efforts like oversampling and augmentation, the VASC class remains the smallest (142 images), resulting in the lowest per-class recall of 0.762.
- **External Validation:** Stage 1's 100% accuracy may stem from the inherent simplicity of binary

classification compared to fine-grained tasks. Independent clinical validation is essential before making any deployment claims.

- **Edge Readiness:** With ~28.6M parameters, the ConvNeXt-Tiny backbone requires knowledge distillation and quantization for real-time edge inference. This work serves as a methodological prototype rather than a production-ready system.

6. Conclusion and Future Enhancements

This paper presented a two-stage deep learning framework for skin lesion analysis that decouples the problem into a binary cancer screening stage and a fine-grained multi-class classification stage, both built on a ConvNeXt-Tiny backbone augmented with Squeeze-and-Excitation and Convolutional Block Attention modules. The framework was trained on a combined Harvard Dataverse and Kaggle corpus using a carefully regularized optimization protocol that includes Focal Loss with inverse-frequency class weighting, MixUp, label smoothing, an annealed dropout schedule, and lesion-aware augmentation strategies such as VASC-Safe CLAHE. On the held-out test sets, the system achieves 99.99% accuracy on the Stage 1 binary screening task and 90.64% accuracy with a macro F1-score of 88.14% on the Stage 2 three-class lesion classification task, with attention-module choices supported by independent ablation evidence from related work (Section 4.3).

Future work will focus on three concrete directions:

- Expanding the lesion taxonomy to include squamous cell carcinoma and actinic keratosis.
- Validating the pipeline on an external clinical dataset to further confirm the robustness of the Stage 1 screener.
- Compressing the model through knowledge distillation and quantization to enable true real-time inference on mobile and point-of-care devices.

It is hoped that the staged formulation, the explicit handling of class imbalance, and the transparent discussion of limitations presented in this work will serve as a useful reference for subsequent dermoscopic classification research.

Acknowledgement

The authors would like to express their sincere gratitude to our supervisor, Er. Nirajan Acharya, Department of Computer Engineering, Institute of Engineering, Tribhuvan University, for their valuable guidance, continuous encouragement, and constructive suggestions throughout the development of the project titled "Skin Cancer Detection and Classification Using Deep Learning." The authors would also like to thank all the faculty members of the Department of Computer Engineering for their constant support and cooperation during the entire project period. Finally, we would like to acknowledge the researchers and authors whose published works, books, and online resources provided the necessary knowledge in the fields of Deep Learning, Convolutional Neural Networks (CNNs), attention mechanisms, ConvNeXt architectures, and medical image analysis, which were essential for the successful completion of this project.

References

[1] W. Limprasert, T. Paipongna, S. Chaowchuen, and S. Vicharueang K. Warin, "Determination of the oral carcinoma and sarcoma in contrast enhanced CT images using deep convolutional neural networks," *Scientific Reports*, vol. 15, no. 1, p. 21672, 2025.

[2] S. H. Cho, J. D. Lee, and H. S. Kim Y. R. Woo, "The human microbiota and skin cancer," *International Journal of Molecular Sciences*, vol. 23, no. 3, p. 1813, 2022.

[3] M. Kumar, V. P. Bhardwaj, S. Kumar, and S. Selvarajan A. Kumar, "A novel skin cancer detection model using modified finch deep CNN classifier model," *Scientific Reports*, vol. 14, no. 1, p. 11235, 2024.

[4] K. Liu, S. Zheng and G. Wang Q. Pan, "A Fine-Grained Image Classification Method Based on ConvNeXt Heatmap Localization and Contrastive Learning," *IEEE Access*, vol. 13, pp. 80123-80132, 2025.

[5] J. Zhang, and X. Wang Z. Wang, "Sheng wu yi xue gong cheng xue za zhi = Journal of biomedical engineering = Shengwu yixue gongchengxue zazhi," *Journal of Biomedical Engineering*, vol. 41, no. 3, pp. 544-551, 2024.