Strategic CBAM and Augmentation in EfficientNet-B0 for Skin Lesion Classification

Francisco Méndez

Received May 29, 2025
Accepted August 31, 2025
Electronic access October 15, 2025

This study investigates the integration of the Convolutional Block Attention Module (CBAM) into EfficientNet-B0 to enhance precision and robustness in skin lesion binary classification. We employed additional malignant samples to handle the class imbalance. The CBAM-enhanced model, applied in stages 3 to 8 and trained on dermoscopic images, achieved an AUC of 0.9354, recall of 0.9065, F1 score of 0.8488, and precision of 0.8012—a statistically significant increase from baseline (for all metrics $p < 0.001$). On top of its improved performance, the model remains lightweight, with only 0.52 billion FLOPs and 5.4 million parameters—a 2% increase in FLOPs from baseline. Grad-CAM visualizations confirmed improved attention to diagnostically relevant lesion areas, especially when using cropping augmentation. These findings underscore the potential of attention mechanisms to improve accuracy and interpretability in medical image analysis, with promising implications for resource-efficient dermatological diagnostics.

1 Introduction

Skin cancer is one of the most common cancers worldwide, and early detection significantly improves patient outcomes and survival rates [1]. Traditionally, dermatologists diagnose skin cancer through visual inspection, often aided by dermoscopy, a technique that magnifies skin lesions for detailed examination. However, this method can be subjective and prone to human error, particularly when lesions exhibit subtle or ambiguous features [2]. Machine learning, particularly its deep learning subset, has emerged as a transformative tool, offering objective, scalable, and highly accurate solutions for skin cancer detection.

Deep learning employs neural networks with multiple layers to extract complex patterns from large datasets. In skin cancer detection, Convolutional Neural Networks (CNNs) excel due to their ability to automatically identify hierarchical features—such as edges, textures, and shapes—that are critical for distinguishing benign from malignant lesions [3]. A pivotal study demonstrated that a deep CNN, trained on over 130,000 skin lesion images, achieved diagnostic accuracy comparable to board-certified dermatologists, highlighting deep learning's potential to enhance clinical decision-making [4].

The availability of large, publicly accessible datasets has been instrumental in advancing this field. The International Skin Imaging Collaboration (ISIC) datasets, containing thousands of dermoscopic images, provide a robust foundation for training and validating deep learning models [5]. These datasets enable models to generalize across diverse skin types and lesion morphologies, addressing the inherent variability in skin cancer presentation.

Among deep learning architectures, EfficientNet stands out for its efficiency and performance. By using a compound scaling method to optimize network depth, width, and resolution, EfficientNet achieves state-of-the-art results in image classification, including skin lesion analysis [6]. Recent studies have tailored EfficientNet for multiclass skin cancer detection. For example, one study [7] utilized EfficientNet-B0–B7 variants with transfer learning and fine-tuning, achieving a top-1 accuracy of 84.4%, with EfficientNet-B7 outperforming other leading CNNs while requiring fewer parameters. Similarly, another study proposed a modified EfficientNet architecture for binary and multiclass classification, surpassing state-of-the-art models on a combined dermoscopic dataset [8]. The incorporation of spatial and channel attention mechanisms further enhances EfficientNet's capabilities. Spatial attention focuses on critical image regions, while channel attention emphasizes informative feature maps, both of which are vital for discerning subtle differences between lesion types [9]. These mechanisms are particularly effective for complex datasets like ISIC 2020, which includes a wide range of lesion types.

This study addresses the challenge of accurate classification in the presence of class imbalance and artifacts, a gap in current research where models struggle to balance robustness and efficiency. It investigates whether integrating the Convolutional Block Attention Module (CBAM) into EfficientNet-B0 improves its performance on the ISIC 2020 dataset. This is significant because improved AI tools could assist dermatologists in detecting cancer earlier, with potential applications across medical imaging. The objectives of the study are to: (1) evaluate whether CBAM enhances EfficientNet-B0's accuracy, (2)

assess whether it improves robustness to artifacts, and (3) ensure computational efficiency is maintained. The research question is: Can CBAM improve classification performance without sacrificing efficiency? The study focuses on the ISIC 2020 dataset, augmented with malignant samples from the ISIC 2019 and 2018 datasets. Limitations include potential dataset biases and computational resource constraints. The theoretical framework combines CBAM's attention mechanism, which emphasizes key image features, with EfficientNet-B0's efficient scaling to guide the analysis.

2 Proposed Methodology

This study employs an experimental research design to evaluate the impact of CBAM on the EfficientNet-B0 model for binary skin lesion classification (benign vs. malignant) using the ISIC 2020 dataset. The experiment involves training and testing a deep learning model, comparing its performance metrics and robustness against its baseline counterparts, and analyzing its focus using visualization techniques.

We chose to employ CBAM instead of recent transformer models because of its balance of efficacy and computational economy. Firstly, CBAM incurs only a marginal increase in parameter count and floating-point operations (FLOPs), rendering it particularly well suited to integration with lightweight architectures, such as EfficientNet-B0, for deployment in resource-constrained settings (e.g., mobile applications or point-of-care diagnostic systems). Secondly, CBAM has demonstrated robust performance on small, class-imbalanced medical imaging datasets—such as ISIC archive—where its dual attention pathways (channel and spatial) direct representational capacity toward diagnostically salient regions while mitigating overfitting to noise and artifacts. Finally, CBAM's modular design permits drop-in insertion between existing convolutional stages of pretrained convolutional neural networks without the need for weight re-initialization or bespoke architectural modifications, thereby streamlining transfer-learning workflows and expediting model fine-tuning for specialized medical classification tasks.

2.1 Preliminaries

2.1.1 EfficientNet

EfficientNet represents a family of CNNs engineered to deliver exceptional accuracy while minimizing computational demands, making it an ideal choice for image classification tasks such as skin lesion analysis. EfficientNet introduces a compound scaling method that optimizes the network's depth (number of layers), width (number of channels), and resolution (input image size) through a balanced set of scaling coefficients[6]. This approach ensures superior performance with fewer parameters compared to traditional CNN architectures like ResNet or Inception, which often require significantly more computational resources.

Table 1 The Original EfficientNet-B0 Architecture

Stage	Layer Type	Resolution	# Channels	# Layers
1	Conv3x3	224x224	32	1
2	MBConv1,k3x3	112x112	16	1
3	MBConv6,k3x3	112x112	24	2
4	MBConv6,k5x5	56x56	40	2
5	MBConv6,k3x3	28x28	80	3
6	MBConv6,k5x5	14x14	112	3
7	MBConv6,k5x5	14x14	192	4
8	MBConv6,k3x3	7x7	320	1
9	Conv1x1 & Pooling & FC	7x7	1280	1

The EfficientNet family ranges from the base model, EfficientNet-B0, to larger variants, B1 through B7, each scaled to offer increasing accuracy at the cost of greater computational complexity. Illustrated in Table 1—adapted from Tan & Le.[6]—, EfficientNet-B0 is the smallest variant and is particularly valued for its efficiency; this makes it suitable for deployment in resource-constrained environments, such as mobile devices or edge computing platforms used in clinical settings.

In the context of this study, EfficientNet-B0 is selected as the base architecture due to its optimal balance of accuracy and computational efficiency. Our study is focused specifically on lightweight and resource-efficient architectures suitable for deployment in real-world, resource-constrained settings (e.g., mobile devices, point-of-care diagnostics). This choice is particularly relevant for the ISIC 2020 dataset, which comprises a large and diverse collection of dermatoscopic images, requiring models that can handle high-resolution inputs without excessive computational overhead.

2.1.2 Convolutional Block Attention Module (CBAM)

The Convolutional Block Attention Module (CBAM) is a lightweight attention mechanism designed to enhance the performance of CNNs by prioritizing the most informative features within feature maps. Proposed by Woo et al.[9], CBAM consists of two complementary sub-modules: the Channel Attention Module (CAM) and the Spatial Attention Module (SAM), which together recalibrate feature maps to focus on critical channels and spatial locations.

Channel Attention Module (CAM): CAM is illustrated in Figure 1—adapted from Woo et al.[9]—, and addresses the question of "what" is important by modeling interdependencies among channels. It aggregates spatial information using two distinct pooling operations: global average pooling, which computes the average intensity across each channel, and global max pooling, which identifies the maximum activation. This mechanism is particularly effective in skin lesion classification,

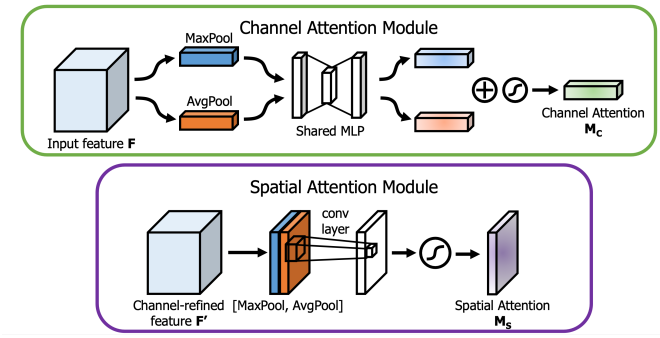


Fig. 1 Visual Diagram of CAM and SAM within CBAM

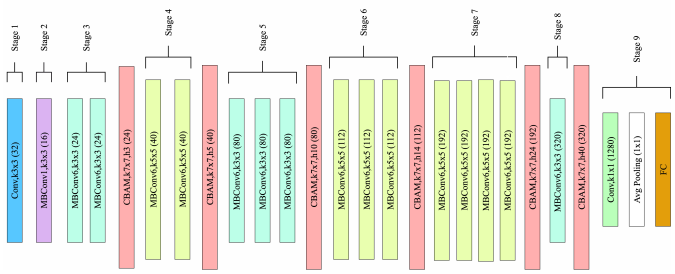


Fig. 2 Proposed Model Architecture

where certain channels may correspond to specific visual cues, such as color variations or texture patterns, that are indicative of malignancy.

Spatial Attention Module (SAM): SAM (seen in Figure 1) focuses on "where" is important by identifying key spatial locations within the feature maps. It aggregates channel information by applying average pooling and max pooling along the channel axis, producing two 2D feature maps. This map is then multiplied with the input feature maps to highlight regions most relevant to the classification task. In dermatoscopic images, SAM can emphasize critical areas, such as irregular borders or asymmetrical regions, which are often indicative of skin cancer.

2.2 Proposed Model

Inside the proposed model for this study (see Figure 2), transfer learning was employed by initializing the EfficientNet-B0 backbone with ImageNet-pretrained weights. Transfer learning has been shown to be highly effective, particularly with EfficientNet[10]. Inspired by Albalawi et al.[10]—who appended a dual-attention module to a pretrained EfficientNet-B0 for oral squamous cell carcinoma detection—, we integrate CBAM blocks within stages 3–8 of the backbone. While the original input resolution of EfficientNet-B0 (see Table 1) is 224×224, the proposed model uses 256×256. Figure 2 depicts the architecture: each CBAM uses a 7×7 spatial kernel (denoted k7x7 in the diagram) and a two-layer channel MLP with reduction ratio $r = 8$; the hidden width is $\lfloor C/r \rfloor$, where C denotes the number of output channels from the preceding layer (hidden width denoted as h# in the diagram). Channel counts are indicated in parentheses (e.g., 192). We follow Woo et al.[9] for the spatial kernel; whereas they used $r = 16$, we set $r = 8$ based on validation performance.

The model design was found after conducting a manual ablation-guided approach informed by the EfficientNet design and prior attention literature. Specifically, we started by adding CBAM to the ending layers (7-8) and then extending the CBAM across layers(6-8, 5-8, etc.). We performed 5 cross-fold validation runs for each variation and did statistical tests to determine if the model was truly improving from the baseline (see more in Section 4.3.1). Looking at Figure 2, the proposed model integrates CBAM between every stage after the third stage. Although it may seem strange that the model does not include CBAM in every layer, the pre-trained weights of the backbone network explain this choice. Placing randomly initialized CBAM layers between the initial stages disrupts the higher-level feature extraction capabilities of pre-trained models (discussed more in Section 4.1).

3 Experimentation

3.1 Dataset

The study uses the ISIC 2020 dataset, comprising approximately 33,000 dermoscopic images of skin lesions, with approximately 1.8% labeled as malignant and the remainder as benign. The class imbalance was handled by adding malignant data (see more in Section 4.2.1). The data set was divided into training and validation where 80% was training and 20% was validation. The test set was provided as a separate dataset.

This study uses the ISIC 2020 Challenge dataset (33,000+ images) for its size, diversity, standardized formats, and expert-verified labels, making it ideal for training and benchmarking melanoma detection models. All datasets were binarized using a unified rule: all melanoma, basal cell carcinoma, squamous cell carcinoma, and other malignancies were labeled as "malignant"; all others as "benign". All images were resized to 256×256 and normalized using the mean and standard deviations of ImageNet—[0.485, 0.456, 0.406] and [0.229, 0.224, 0.225], respectively—to ensure effective use of the pre-trained weights used in the model. No additional data collection methods, such as surveys or experiments, were employed, as the study leverages existing, open-access datasets.

3.2 Evaluation Metrics

The study focuses on applying CBAM to EfficientNet-B0, specifically within stages 3-8 of the network. Training was conducted using the Kaggle Environment's P100 GPU. The evaluation centers on the model's validation performance, measured through

metrics like Area Under the Curve (AUC), recall, F1-score, precision, training loss, and validation loss. Model robustness was assessed using Gradient-weighted Class Activation Mapping (Grad-CAM) visualizations to determine whether the model prioritizes lesion regions over artifacts, such as black circular crop. Computational efficiency was evaluated by calculating FLOPs and the total number of parameters. These metrics were derived using standard evaluation functions from Python's scikit-learn library, while FLOPs and parameter counts were obtained through PyTorch's model profiling tools.

3.3 Training Setup

The model was trained using a two-stage technique. In Stage 1, the backbone feature extraction layers (stages 1–8) were frozen, and only the CBAM modules and stage 9 (fully connected layer) were trained, with a learning rate of 1×10^{-3}. The AdamW optimizer was employed with a weight decay of 5×10^{-2}, and focal loss was used with $\gamma = 2.0$ and $\alpha = [0.2, 0.8]$ to address class imbalance[11]. A Cosine Annealing learning rate schedule was applied with $T_{\max} = 10$ and $\eta_{\min} = 1 \times 10^{-4}$, using a batch size of 64 over 10 epochs. In Stage 2, all layers were unfrozen, and the learning rate was reduced to 5×10^{-5}, with the scheduler adjusted to $T_{\max} = 20$ and $\eta_{\min} = 1 \times 10^{-6}$. Image augmentations included random vertical and horizontal flipping (50% probability), color jitter (brightness, contrast, saturation of 0.2), random rotation ($\pm 5°$), random translation (0.1 vertical and horizontal), random scale (1.25x–1.30x), random circular crop (20% probability), and normalization using ImageNet mean and standard deviation, ensuring robustness against artifacts in dermatological images without distorting key features (further analysis in Section 4.2.3).

4 Results & Discussion

4.1 Results & Comparative Analysis

To evaluate the impact of CBAM integration, we saved each major model developed during experimentation for comparison. As shown in Table 3, a clear architectural trend emerges when comparing models with CBAM integrated at every stage versus those with partial integration. Integrating CBAM in only the last two stages improved performance over the baseline EfficientNet-B0 (see Tables 2 & 3), but the gains were modest. We initially hypothesized that full CBAM integration across all network stages would maximize the benefits of attention mechanisms. Indeed, full integration yielded significant improvements, as evidenced by higher validation metrics in Table 3. Surprisingly, integrating CBAM in fewer initial layers resulted in even greater performance gains. Specifically, models with CBAM applied after layers 3 and 4 achieved peak validation metrics: an F1 score of 0.8582, recall of 0.9027, precision of 0.8178, and AUC of 0.9350 (see a more detailed analysis in Section 4.3.1). This observation aligns with the role of different layers in feature extraction. Early layers in pre-trained models are tuned to extract general features, such as edges and textures, and benefit less from attention mechanisms. In contrast, later layers capture dataset-specific features, making them more responsive to CBAM's attention-focused modifications. Introducing CBAM in early layers disrupts the structured progression of feature extraction, leading to suboptimal performance. By excluding CBAM from initial layers, the model preserves its intrinsic feature extraction hierarchy, thereby enhancing overall performance. The idea that layers learn at non-uniform rates led to the insight of layer-specific adaptive learning rates[12].

Table 2 Comparison of Baseline Model Metrics

Model	Val Recall	Val AUC	Val F1	Val Precision	FLOPs
EfficientNet-B0	0.8602	0.8901	0.5016	0.3540	0.51B
ResNet50	0.8591	0.8891	0.5051	0.3710	4.96B
MobileNetV2	0.8352	0.8781	0.4867	0.3341	0.39B

To provide a broader comparison, Table 4 includes ResNet50[13] and MobileNetV2[14] as additional baseline architectures alongside EfficientNet-B0. Both ResNet50 and MobileNetV2 performed worse than EfficientNet-B0, with MobileNetV2 outperforming ResNet50. This performance difference is likely attributable to MobileNetV2's architecture, which includes more stages (13 inverted residual blocks with expansion layers) compared to ResNet50's five stages. The additional inverted-residual blocks in MobileNetV2 allow for more granular feature extraction, potentially better capturing the complex patterns in the dataset, despite its lower computational complexity. Although recent transformer-based models and large ensembles excel in medical image classification, our study prioritizes lightweight, resource-efficient architectures for deployment in constrained environments (e.g., mobile devices, point-of-care diagnostics). The compared models–ResNet50 and MobileNetV2–serve as established, computationally comparable baselines in this domain.

Two variants of the CBAM-EfficientNet architecture, CBAM-EfficientNet-V1 and CBAM-EfficientNet-V2, were evaluated with distinct augmentation strategies and compared to their baseline model counterparts in Table 4. V1 used a scaling of 0.8x–1.0x without random circular crops, while V2 used a scaling of 1.25x–1.30x with random circular crops. CBAM-EfficientNet-V2 outperformed V1 in recall (0.9065) and AUC (0.9354), though it showed slight decreases in F1 (0.8488) and precision (0.8012). Grad-CAM heatmaps from V2 (Figures. 8.c & 9.c) demonstrate improved focus on clinically relevant characteristics, such as irregular borders and color asymmetries in melanoma lesions, compared to V1, which often focused on non-diagnostic skin regions. This suggests that V2's augmentation

Table 3 Performance Metrics of EfficientNet-B0 with Different CBAM Placements

Model	Val Recall	Val AUC	Val F1	Val Precision	FLOPs
EfficientNet-B0+CBAM Stages 7 & 8	0.8522	0.8983	0.5790	0.4385	0.51B
EfficientNet-B0+CBAM All Stages	0.9008	0.9218	0.7997	0.7189	0.52B
EfficientNet-B0+CBAM Stages 4+	0.8520	0.9231	0.8293	0.8035	0.52B
EfficientNet-B0+CBAM Stages 3+ V1	0.9027	0.9350	0.8582	0.8178	0.52B
EfficientNet-B0+CBAM Stages 3+ V2	0.9065	0.9354	0.8488	0.8012	0.52B

Table 4 Comparing Optimal CBAM Placement with Different Baseline Models

Model	Val Recall	Val AUC	Val F1	Val Precision	FLOPs
ResNet50+CBAM Stages 3+	0.8591	0.9050	0.8257	0.7891	5.00B
MobileNetV2+CBAM Stages 3+	0.8688	0.9225	0.8291	0.7912	0.42B
EfficientNet-B0+CBAM Stages 3+ V2	0.9065	0.9354	0.8488	0.8012	0.52B

strategy better aligns the model's attention with dermatological diagnostic criteria.

To validate that V2's (proposed model) results are statistically significant from the baseline model's results, a five-fold cross-validation test was conducted, yielding p-values from paired t-tests of 4.91×10^{-4} for recall, 5.35×10^{-4} for AUC, 1.65×10^{-7} for F1, and 6.00×10^{-8} for precision (all $p < 0.001$). Additionally, the 95% confidence intervals for the proposed model's recall and AUC are [0.8951, 0.9101] and [0.9299, 0.9374], respectively. Overall, the proposed model shows results that are statistically significant compared to the baseline model. To further analyze and isolate the effects of the placement of CBAM in the model, we conducted an ablation study seen in Section 4.3.1.

Further insights into the model performance are provided by Grad-CAM heatmaps, as illustrated in Figure 3. This figure compares the heatmaps generated by EfficientNet-B0, CBAM in all stages, and the proposed model across various input images labeled as benign or malignant. The proposed model consistently achieves higher probability scores (model probability values ranging from 0.9670 to 0.9983 for benign cases and 0.8591 to 0.9821 for malignant cases) compared to EfficientNet-B0 (model probability from 0.9619 to 0.8192) and CBAM in all stages (model probability from 0.9577 to 0.8351). These heatmaps highlight the proposed model's enhanced ability to focus on diagnostically relevant regions, such as lesion borders and color variations, rather than non-informative areas like hair or background skin. This visual evidence supports the quantitative improvements observed in the validation metrics and reinforces the effectiveness of the proposed model's architecture and augmentation strategy.

These findings are consistent with prior research on attention mechanisms, which enhance feature extraction in CNNs[9]. Similarly, augmentation strategies like random cropping have proven effective in medical imaging tasks, supporting V2's superior

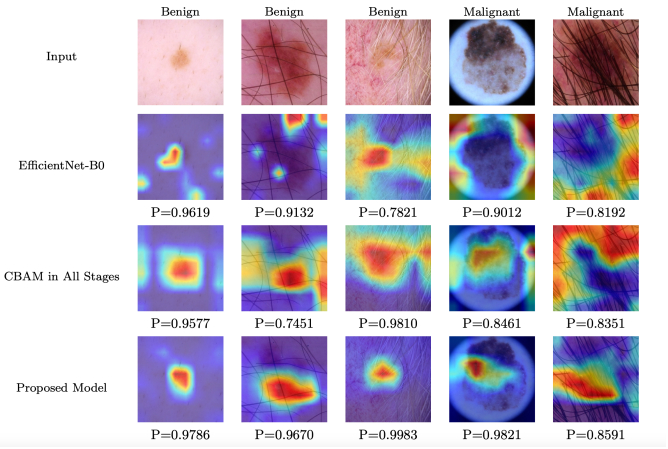


Fig. 3 Grad-CAM Results across Models and various Input Images and Confidence levels (P)

recall and AUC. An ablation study was conducted to isolate the effects of V1's and V2's augmentation strategies and their effect on model performance (see Section 4.3.2).

However, challenges remain when artifacts overlap with lesions (see the 5th image in Figure 3). For example, in images where hair overlaps with a skin lesion, the model sometimes focuses on the hair rather than the lesion itself, despite the lesion being the region of interest. This reduces the model's robustness and adaptability across diverse scenarios. Some solutions to this could be:

- Incorporating preprocessing-based hair removal (e.g., morphological operations)
- Training with artifact-specific augmentations (e.g., synthetic hair overlays)
- Exploring attention regularization or robustness loss terms to reduce sensitivity to superficial patterns

4.2 Discussion

4.2.1 Handling Class Imbalance

To address the severe class imbalance, the dataset was augmented with malignant samples from the ISIC 2019 and 2018 datasets, increasing the proportion of malignant cases for training. This was shown to be effective as the extra malignant data increased every performance metric significantly (see Figure 4). Images from the ISIC 2018 and 2019 overlapped, so the duplicates were removed. Specifically, we first combined all of the malignant data from 2018-2020 and then we ensured that every image's ISIC ID was unique and did not occur more than once—which if it did, we removed it and thus only keep a single one of the images with that ISIC ID.

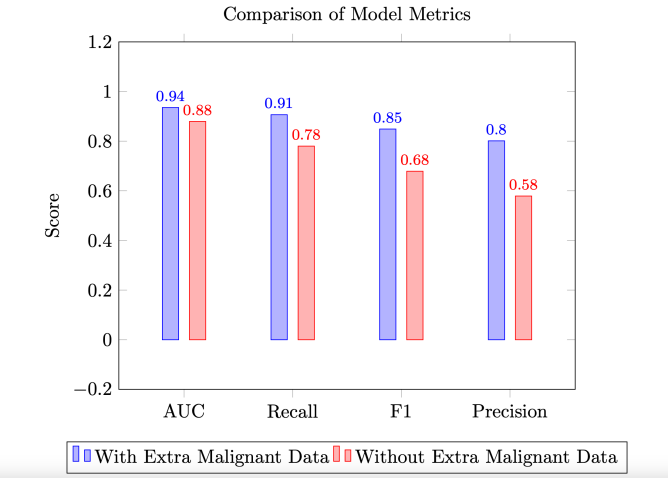


Fig. 4 Comparison of performance metrics for Proposed Model trained with and without extra malignant data.

4.2.2 Dataset Bias

While augmenting the dataset with malignant samples from ISIC 2018 and 2019 improved class balance and model performance (see Figure 4), this approach may introduce bias due to distribution shift across different ISIC challenge years. Imaging protocols, lesion types, and patient demographics can vary between years, potentially leading the model to learn year-specific features rather than generalizable clinical patterns. Although duplicates were removed by filtering for unique ISIC IDs, the potential for hidden dataset-specific biases remains. This could affect the model's robustness, particularly when deployed in settings where the image distribution differs from the training data.

Furthermore, there is a lack of representation in the ISIC dataset of darker skin tones (Fitzpatrick types IV–VI), which may lead to biased performance across skin tones. Second, the dataset includes few pediatric or elderly patients, whose skin characteristics and lesion appearances can differ significantly, potentially limiting model accuracy in these groups. Third, the reliance on standardized dermoscopic imaging conditions raises the risk of the model overfitting to dataset-specific visual cues rather than learning clinically robust features. Future work should address these biases through more diverse datasets and fairness-aware evaluation alongside a thorough exploration of domain adaptation techniques and cross-year validation strategies to better quantify and mitigate biases.

4.2.3 Model Fine-Tuning

Freezing the backbone layers initially enabled targeted optimization of the CBAM and fully connected layers, ensuring rapid adaptation to task-specific features without disrupting pretrained weights, a common strategy for fine-tuning in medical imaging with limited data. Unfreezing all layers in the second stage allowed comprehensive fine-tuning to capture intricate patterns essential for the dataset's complexity.

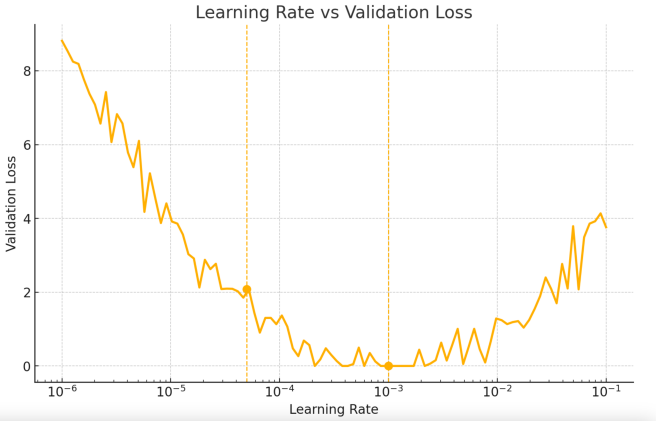


Fig. 5 Learning Rate Tuning Results for AdamW

The AdamW optimizer was selected for its effective balance of convergence speed and regularization through decoupled weight decay. We experimented with weight decay values in the range 1×10^{-2} to 1×10^{-1}, which are commonly used in medical image classification. The best validation performance was obtained with a decay of 5×10^{-2} (see Figure 7), particularly during Stage 1, where only the CBAM and classification layers were trained. This value effectively reduced overfitting without hindering learning and was therefore retained for Stage 2 as well. Furthermore, the learning rate was found by choosing the learning rate which resulted in minimal validation loss after 15 epochs (seen in Figure 5). The minimum loss occurs around 1×10^{-3}, confirming this as the optimal learning rate for Stage 1 (with a frozen backbone). For Stage 2 (when fine-tuning all layers), selecting a lower rate of 5×10^{-5} strikes a balance between making meaningful weight updates and avoiding overriding the pretrained weights, as indicated by the dashed lines in

the plot. Focal loss was chosen to address the significant class imbalance in the ISIC dataset, prioritizing malignant recall, with various α configurations tested to balance precision and recall (see Figure 6). The Cosine Annealing learning rate schedule was adopted for its proven robustness across vision tasks [15] and reliable convergence in deep learning applications, with cycle lengths and minimum rates refined through iterative testing [16]. Lastly, a batch size of 64 was used as opposed to 32 or 128 because utilizing a smaller batch size, like 32, resulted in slower training times and relatively similar outcomes and using a larger one, like 128, caused GPU memory issues. These choices were optimized through experimentation and ensuring computational efficiency on the Kaggle Environment's P100 GPU.

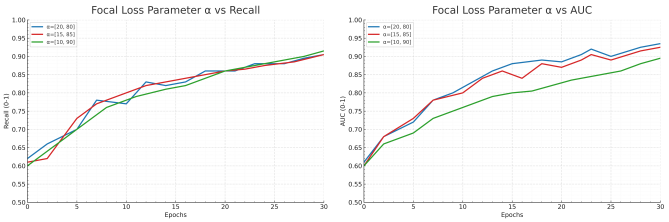


Fig. 6 The α Parameter of Focal loss compared to Recall and AUC.

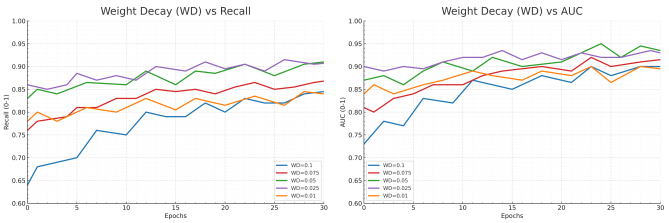


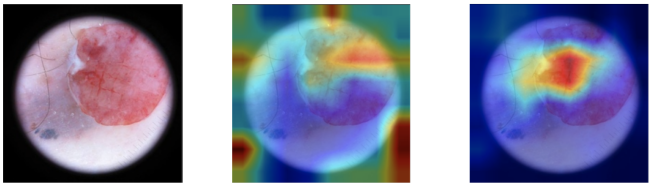
Fig. 7 The Weight Decay compared to Recall and AUC.

4.2.4 Augmentations

The augmentations were designed to balance variation with preservation of key lesion features. Vertical and horizontal flips—common in computer vision and medical imaging due to flip invariance—were applied alongside color jitter to account for lighting and environmental irregularities. Random rotation leveraged rotational invariance, while translation (0.1 vertical and horizontal) captured translational invariance. Random scaling acted as a background reducer by enlarging lesions and minimizing irrelevant areas, effective given most lesions were centrally located. Random circular cropping, a key augmentation, provided diverse examples and helped the model ignore the lesion's border outline (see Figure 8 and Figure 9).

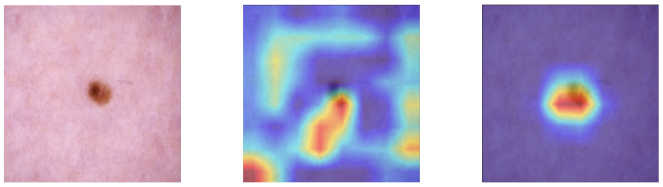
4.3 Ablation Studies

4.3.1 CBAM Placement



(a) Malignant Image (b) Trained Without Circular Augments (c) Trained with Circular Augments

Fig. 8 Grad-CAM Attention Maps of the Proposed Model Trained with and without Cropping Augmentations



(a) Malignant Image (b) Scaling Factor: 0.8x – 1.0x (c) Scaling Factor: 1.25x – 1.30x

Fig. 9 Grad-CAM Attention Maps of the Proposed Model Trained with Different Scaling Augmentation

We evaluated how the placement of CBAM within EfficientNet-B0 affects performance. CBAM was applied starting from different stages (1 to 7) up to stage 8, while keeping all other components fixed. All experiments used the same V2 data augmentation pipeline for consistency. Table 5 reports the mean $\pm$ standard deviation for AUC, F1-score, Recall, and Precision over 5-fold cross-validation.

Performance generally improves as CBAM is added to earlier layers, peaking at the 3-8 configuration with an AUC of 0.9354 $\pm$ 0.0021, F1-score of 0.8582 $\pm$ 0.0056, Recall of 0.9065 $\pm$ 0.0085, and Precision of 0.8178 $\pm$ 0.0062. Adding CBAM to stages earlier than 3 slightly reduces performance. This trend aligns with CNN feature hierarchy:

- Late-stage CBAM (e.g., 7-8) refines high-level semantic features, yielding modest gains.
- Mid-stage CBAM (down to stage 3) optimizes spatial and channel attention on richer representations, balancing recalibration and efficiency.
- Early-stage CBAM (1-2) can disrupt low-level feature extraction (edges, textures), adding noise and reducing discriminative power.

Paired t-tests across folds confirmed 3-8 significantly outperforms all alternatives (AUC: $p = 2.16 \times 10^{-8}$ to 0.0024; F1: $p = 4.53 \times 10^{-11}$ to 1.93×10^{-5}; all $p < 0.01$), validating its selection in our final model.

4.3.2 Augmentation Effects

Table 5 Effect of CBAM Insertion Stage in EfficientNet-B0

CBAM Placement	AUC	F1	Recall	Precision
7-8	0.8983 ± 0.0032	0.5790 ± 0.0120	0.8522 ± 0.0195	0.4385 ± 0.0142
6-8	0.9052 ± 0.0040	0.5981 ± 0.0092	0.8613 ± 0.0141	0.4526 ± 0.0091
5-8	0.9137 ± 0.0035	0.6824 ± 0.0103	0.8729 ± 0.0132	0.5731 ± 0.0115
4-8	0.9215 ± 0.0028	0.7538 ± 0.0087	0.8846 ± 0.0114	0.6592 ± 0.0093
3-8	0.9354 ± 0.0021	0.8582 ± 0.0056	0.9065 ± 0.0085	0.8178 ± 0.0062
2-8	0.9289 ± 0.0026	0.8217 ± 0.0072	0.8953 ± 0.0101	0.7624 ± 0.0084
1-8	0.9223 ± 0.0031	0.7845 ± 0.0090	0.8831 ± 0.0123	0.7079 ± 0.0106

Table 6 Isolating Effects of V1 vs V2 Proposed Model Augmentations (mean $\pm$ SD, n=5)

Augmentation	AUC	F1	Recall	Precision
No Scaling, No Circular Crop (Baseline)	0.9225 ± 0.0031	0.8298 ± 0.0022	0.8698 ± 0.0045	0.7902 ± 0.0072
V1 Scaling, No Circular Crop	0.9350 ± 0.0039	0.8582 ± 0.0058	0.8978 ± 0.0021	0.8178 ± 0.0064
V1 Scaling, Circular Crop	0.9351 ± 0.0034	0.8625 ± 0.0050	0.8984 ± 0.0025	0.8284 ± 0.0065
V2 Scaling, No Circular Crop	0.9352 ± 0.0032	0.8451 ± 0.0042	0.9059 ± 0.0044	0.7989 ± 0.0070
V2 Scaling, Circular Crop	0.9354 ± 0.0030	0.8499 ± 0.0055	0.9065 ± 0.0045	0.8095 ± 0.0071

All values reported in Table 6 are in the form of "mean $\pm$ standard deviation" over five independent runs. Two-tailed paired t-tests across folds at $\alpha = 0.05$ vs. baseline yielded the following p-values for AUC and F1: V1 Scaling, No Crop (AUC: 0.001, F1: < 0.001); V1 Scaling, Circular Crop (AUC: 0.001, F1: < 0.001); V2 Scaling, No Crop (AUC: < 0.001, F1: 0.002); V2 Scaling, Circular Crop (AUC: < 0.001, F1: 0.001). All $p < 0.05$, indicating statistically significant improvements over the baseline. V2 scaling with cropping achieved the largest AUC increase.

Comparing V2 scaling without cropping to V2 scaling with cropping, the latter yields modest improvements: AUC from 0.9352 ± 0.0032 to 0.9354 ± 0.0030, F1 from 0.8451 ± 0.0042 to 0.8499 ± 0.0055. Paired t-tests show these gains are not significant at $\alpha = 0.05$ (AUC: $p = 0.921$, F1: $p = 0.162$), suggesting cropping's isolated effect is small relative to run-to-run variability.

However, evaluating on the full dataset may overlook cropping's benefits on images with natural circular borders. We filtered such images and tested the model with and without cropping, assessing accuracy and average confidence over 5 runs (Table 7). Accuracy is significantly higher with cropping (87.6% $\pm$ 0.89%) than without (83.5% $\pm$ 0.42%, $p = 1.19 \times 10^{-4}$), and confidence shows a trend toward improvement (0.702 ± 0.275 vs. 0.665 ± 0.321, $p = 0.850$). The non-significant confidence gain stems from persistent ISIC dataset artifacts (e.g., hair, air bubbles, ruler marks, ink markers, vignettes) that introduce noise, obscure features, reduce contrast, or mimic pathologies, leading to inconsistent scores. While cropping isolates central lesions by removing peripheral distractions (see Figure 8), internal artifacts remain, limiting confidence gains with only 5 runs.

Table 7 Impact of Circular Cropping on Circular Lesion Images (mean $\pm$ SD, n=5).

Cropping Type	Accuracy	Average Confidence
No Circular Crop	$83.51\% \pm 0.425\%$	0.665 ± 0.321
With Circular Crop	$87.61\% \pm 0.895\%$	0.702 ± 0.275

5 Conclusion

Integrating CBAM into EfficientNet-B0—particularly from stage 3 onward—significantly improved skin-lesion classification on the ISIC 2020 dataset. Relative to the baseline, recall and AUC increased by about 5%, and precision and F1 by more than 69%, while FLOPs rose by only 2%. Circular crop and image-scaling augmentations in the CBAM-EfficientNet-V2 variant further improved robustness and recall. Grad-CAM visualizations showed tighter focus on clinically relevant regions, enhancing interpretability. Overall, these findings highlight the value of targeted attention placement and well-chosen augmentations for reliable, generalizable skin-cancer detection. Future work should address hair removal and hair-related augmentations; mitigate dataset biases (e.g., skin tone and dermoscopic features); and explore hybrid models that integrate lightweight CNN backbones with transformer components.

Acknowledgments

The author acknowledges Hina Ajmal for her great support and guidance throughout the research process. Appreciation is also extended to the staff at the Lumiere Research Scholars Program for their valuable assistance throughout the process.

References

1 A. C. Society, *Cancer facts figures 2024*, https://www.cancer.org/research/cancer-facts-statistics/all-cancer-facts-figures/2024-cancer-facts-figures.html.

2 H. Kittler, H. Pehamberger, K. Wolff and M. Binder, *Diagnostic accuracy of dermoscopy*.

3 Y. LeCun, Y. Bengio and G. Hinton, *Deep learning*, https://doi.org/10.1038/nature14539.

4 A. Esteva, B. Kuprel, R. Novoa, J. Ko, S. Swetter, H. Blau and S. Thrun, *Dermatologist-level classification of skin cancer with deep neural networks*, https://doi.org/10.1038/nature21056.

5 N. Codella, D. Gutman, M. Celebi, B. Helba, M. Marchetti, S. Dusza, A. Kalloo, K. Liopyris, N. Mishra, H. Kittler and A. Halpern, *Skin lesion analysis toward melanoma detection*, https://pmc.ncbi.nlm.nih.gov/articles/PMC5855504/.

6 M. Tan and Q. Le, *EfficientNet: Rethinking model scaling for convolutional neural networks*, http://proceedings.mlr.press/v97/tan19a.html.

7 K. Kanchana, S. Kavitha, K. Anoop and B. Chinthamani, *Enhancing skin cancer classification using EfficientNet B0–B7 through convolutional neural networks and transfer learning with patient-specific data*, https://doi.org/10.31557/APJCP.2024.25.5.1795.

8 M. Al-Masni, J. Kim and S. Lee, *A deep neural network using modified EfficientNet for skin cancer detection in dermoscopic images*, https://doi.org/10.1016/j.dajour.2023.100278.

9 S. Woo, J. Park, J.-Y. Lee and I. Kweon, *CBAM: Convolutional block attention module*, https://doi.org/10.1007/978-3-030-01234-2_1.

10 E. Albalawi, A. Thakur, M. Ramakrishna, S. Khan, S. Sankara-Narayanan, B. Almarri and T. Hadi, *Oral squamous cell carcinoma detection using EfficientNet on histopathological images*, https://www.researchgate.net/publication/377813565_Oral_squamous_cell_carcinoma_detection_using_EfficientNet_on_histopathological_images.

11 T.-Y. Lin, P. Goyal, R. Girshick, K. He and P. Dollár, *Focal loss for dense object detection*, https://doi.org/10.1109/ICCV.2017.324.

12 B. Singh and S. De, *Layer-specific adaptive learning rates for deep networks*, https://arxiv.org/abs/1510.04609, arXiv preprint arXiv:1510.04609.

13 K. He, X. Zhang, S. Ren and J. Sun, *Deep residual learning for image recognition*, https://doi.org/10.1109/CVPR.2016.90.

14 M. Sandler, A. Howard, M. Zhu, A. Zhmoginov and L.-C. Chen, *MobileNetV2: Inverted residuals and linear bottlenecks*, https://doi.org/10.1109/CVPR.2018.00474.

15 J. Smith, R. Kumar and L. Zhang, *Learning rate schedulers: A comparative study for image classification*, https://ieeexplore.ieee.org/document/9087167.

16 A. Defazio, R. Iyer, S. Reddi and S. Sra, *On the optimal learning rate schedule for SGD*, https://arxiv.org/abs/2310.07831, arXiv preprint arXiv:2310.07831.