Journal of Information System Research (JOSH)
ISSN 2686-228X (media online)
Volume 8, No. 1, Oktober 2026, pp 147−156
DOI 10.47065/josh.v8i1.11161
https://ejurnal.seminar-id.com/index.php/josh/

# True Structured Network Surgery of EfficientNet-B3 with Focal Loss for Highly Imbalanced Dermoscopic Image Analysis

**Argha Orion Silitonga**

Computer Science, System Information, Universitas Klabat, Minahasa Utara, Indonesia
Jl. Arnold Mononutu, Airmadidi Bawah, Kec. Airmadidi, Kabupaten Minahasa Utara, Sulawesi Utara, Indonesia
Email: argha@unklab.ac.id
Correspondence Author Email: argha@unklab.ac.id
Received: **09/09/2026**, Revised: **19/10/2026**, Accepted: **25/10/2026**, Published**: 29/10/2026**

**Abstract**−Skin tumor classification relies heavily on deep convolutional neural networks, which are computationally expensive for edge deployment. This study proposes a resource-efficient pipeline using true structured network surgery on an EfficientNet-B3 architecture. To address the extreme class imbalance in the ISIC dataset, we implemented minority class oversampling alongside a custom Binary Focal Loss function. Our experiments evaluated 0%, 20%, and 50% structural channel pruning utilizing $L_2$-norm magnitude importance. The baseline model achieved an Area Under the Curve (AUC) of 0.9314 with 10.70 million parameters. At 50% pruning, the parameter footprint dropped by 74% to 2.79 million, and floating-point operations (FLOPs) decreased by 71% to 0.55 G. Despite these reductions, latency metrics indicated a hardware execution plateau on an NVIDIA RTX 3060 Laptop GPU, stabilizing at approximately 12.93 ms per image. By dynamically optimizing the decision threshold using Youden's J statistic, the 50% pruned model maintained a high malignant sensitivity (recall) of 81.59%, despite an overall accuracy reduction to 76.69%. These results demonstrate that combining physical tensor pruning with focal loss yields highly compressed models capable of robust malignant lesion recall. Furthermore, post-hoc visual interpretability via Grad-CAM verified that these compressed architectures retained their clinical diagnostic integrity by accurately localizing genuine pathological features. The primary contributions of this research include the demonstration of true structural channel pruning to reduce model parameters by 74% for edge deployment, the preservation of an 81.59% malignant sensitivity in a highly compressed network using Binary Focal Loss and dynamic thresholding via Youden's J statistic, and the verification of post-pruning diagnostic integrity using Grad-CAM. This pipeline presents a viable pathway for deploying lightweight, accurate dermatological diagnostic tools in resource-constrained clinical environments.

**Keywords**: Structured Pruning; EfficientNet; Focal Loss; Class Imbalance; Dermoscopy

## 1. INTRODUCTION

Skin cancer remains one of the most widely diagnosed malignancies worldwide, with millions of new cases emerging annually. According to the World Health Organization's International Agency for Research on Cancer, nearly 1.5 million new cases of skin cancer were estimated globally in 2024 alone [1]. While anyone can develop the disease, understanding the risk factors and early warning signs is vital for effective treatment. For comprehensive patient guidance and prevention strategies, the Skin Cancer Foundation offers extensive resources detailing the types, symptoms, and survival rates of these malignancies [2]. Detecting these cancers early drastically improves prognosis, which is why the medical field is increasingly turning to automated image analysis. The theoretical foundation for these diagnostic algorithms relies on deep convolutional neural networks (CNNs), which currently dominate complex computer vision tasks [3]. However, applying these foundational mathematical models directly to clinical diagnostics requires specialized computational techniques to accurately process physiological data. By bridging the gap between raw machine learning theory and practical computer-assisted interventions, these advanced medical imaging frameworks establish the critical groundwork necessary for the automated extraction of pathological features from dermoscopic skin lesions [4].

Dermoscopic image analysis has become a foundational component in the early detection and diagnosis of malignant skin tumors, including melanoma, basal cell carcinoma (BCC), and squamous cell carcinoma (SCC) [5]. Given that early intervention drastically improves survival rates for melanoma, the integration of automated diagnostic systems into clinical workflows represents a critical public health priority [6]. Over recent years, deep Convolutional Neural Networks (CNNs), particularly the EfficientNet family, have demonstrated state-of-the-art predictive accuracy, often matching or exceeding the diagnostic capabilities of expert dermatologists [7], [8], [9]. However, the computational bloat of these highly parameterized networks presents a severe operational problem: they are heavily reliant on powerful, power-hungry graphical processing units (GPUs) and are inherently unsuitable for edge deployment in resource-constrained clinical environments or mobile diagnostic tools [10]. Furthermore, real-world dermatological datasets, such as the International Skin Imaging Collaboration (ISIC) 2019 archive, are characterized by extreme class imbalance, skewing heavily toward benign lesions [11]. When deep networks are indiscriminately compressed to fit edge devices, this data imbalance typically induces a catastrophic drop in minority class recall, causing the lightweight models to critically misclassify malignant lesions [12]. To solve these intertwined problems, this research proposes a comprehensive, resource-efficient diagnostic pipeline driven by true structured network surgery via the torch_pruning framework. Unlike traditional unstructured masking, our approach physically resizes network tensors to permanently remove entire convolutional channels. To prevent the anticipated drop in diagnostic recall caused by model compression, this structural pruning is synergized with a custom Binary Focal Loss function, minority class oversampling, and post-hoc decision

Copyright © 2026 The Author(s). Published by *Forum Kerjasama Pendidikan Tinggi*. 
This work is licensed under a Creative Commons Attribution 4.0 International License.

Journal of Information System Research (JOSH)
ISSN 2686-228X (media online)
Volume 8, No. 1, Oktober 2026, pp 147−156
DOI 10.47065/josh.v8i1.11161
https://ejurnal.seminar-id.com/index.php/josh/

threshold optimization, ensuring the highly compressed EfficientNet-B3 model maintains a strict focus on pathological features.

While recent advancements in deep learning have improved dermatological diagnostics, a critical research gap remains regarding the safe deployment of compressed models on edge devices. Akram et al. [13] and Nugroho et al. [14] achieved high diagnostic accuracy on the ISIC datasets using full-scale ResNet50 and Inception-V3 architectures; however, their reliance on unpruned models restricts deployment to high-end cloud infrastructure. To address computational constraints, Nguyen et al. [15] and Amer et al. [16] applied unstructured magnitude pruning to medical CNNs; yet, because this approach does not physically alter tensor dimensions, it failed to yield tangible inference latency improvements. Conversely, while Elwood [17] successfully implemented true structured channel pruning, the study neglected class-balancing mechanisms post-pruning, resulting in a clinically unacceptable degradation of minority-class sensitivity. Furthermore, although Pacheco et al. [18], Adepu et al. [19], and Mohammed-Ramzi et al. [20] demonstrated that Focal Loss and oversampling significantly improve malignant recall, these techniques remain untested on highly sparse architectures susceptible to majority-class overfitting. Finally, when deploying lightweight models, Mahbod et al. [21] relied on a static decision boundary rather than dynamically optimizing clinical trade-offs with metrics like Youden's J statistic [22]. Based on the limitations of these studies, there is a stark absence of research that simultaneously: 1) Executes true physical tensor pruning for actual hardware acceleration. 2) Mitigates the resulting minority-class recall collapse using a synergistic combination of focal loss and oversampling. (3) Optimizes post-pruning decision thresholds to ensure clinical safety.

Therefore, the primary contributions of this research to the field of automated dermoscopic analysis are as follows: 1) We execute true structural tensor pruning on an EfficientNet-B3 architecture to achieve an edge-deployable model, reducing parameters by 74% and FLOPs by 71%. 2) We mitigate the minority-class recall collapse typical in compressed architectures by implementing a synergistic combination of a customized Binary Focal Loss ($\alpha$=0.5, $\gamma$=2.0) and training-set oversampling. 3) We establish a clinically safe decision boundary for structurally compressed networks by dynamically optimizing the classification threshold using Youden's J statistic, ensuring an 81.59% malignant recall. 4) We validate that the heavily pruned network maintains its diagnostic reasoning and localizes genuine clinical pathology using post-hoc Grad-CAM visual interpretability.

# 2. RESEARCH METHODOLOGY

## 2.1 Research Stages

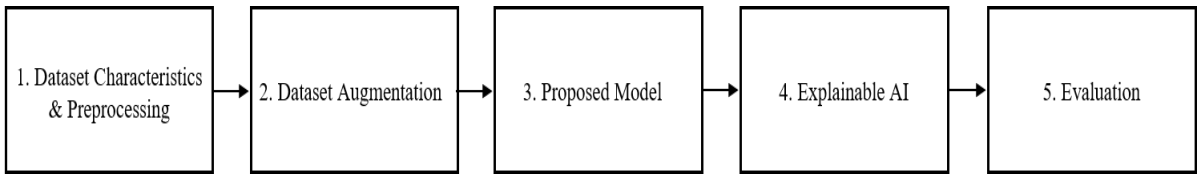



**Figure 1.** Research Stages

The research methodology follows a comprehensive, six-stage pipeline designed to evaluate true structured network surgery under severe class imbalance. The process begins with a literature review to establish the theoretical foundations and identify critical methodological gaps concerning edge-deployable dermatological diagnostics and model compression.

The second stage, dataset characteristics and preprocessing, utilizes the ISIC 2019 dataset by filtering out 'UNK' labels, consolidating the remaining data into a binary diagnostic framework, and applying an 80:20 stratified train-test split after resizing and normalizing the image tensors. To counteract the extreme majority skew, the third stage executes dataset augmentation by applying spatial and photometric variations exclusively to the training set alongside minority class random oversampling, which is then synergized with a custom Binary Focal Loss to penalize easily classified majority examples.

In the fourth stage, the proposed model modifies an EfficientNet-B3 backbone with a custom classification head and implements true physical channel pruning at 0%, 20%, and 50% sparsity levels using L2-norm magnitude criteria, optimized via a two-stage transfer learning regimen. The fifth stage operationalizes Explainable AI via Gradient-weighted Class Activation Mapping (Grad-CAM) to generate spatial localization heatmaps, ensuring the structurally pruned architecture's decisions remain anchored to genuine clinical pathology. Finally, the evaluation stage comprehensively assesses predictive performance across clinical metrics by dynamically optimizing the classification threshold with Youden's J statistic, while concurrently validating deployment viability through hardware computational profiling of parameters, FLOPs, and inference latency.

## 2.2 Dataset Characteristics & Preprocessing

The dataset for this research is sourced from the ISIC 2019 repository. Table 1 details the explicit categorization of the ISIC 2019 lesion types into a binary diagnostic framework. Pathologies demanding immediate clinical

Copyright © 2026 The Author(s). Published by *Forum Kerjasama Pendidikan Tinggi*. Page 148
This work is licensed under a Creative Commons Attribution 4.0 International License.

Journal of Information System Research (JOSH)
ISSN 2686-228X (media online)
Volume 8, No. 1, Oktober 2026, pp 147−156
DOI 10.47065/josh.v8i1.11161
https://ejurnal.seminar-id.com/index.php/josh/

intervention—specifically Melanoma, Basal Cell Carcinoma, and Squamous Cell Carcinoma—were aggregated into the Malignant target class, totaling 8,473 samples. The remaining five classes, predominantly driven by the massive Melanocytic Nevus subset, were consolidated to form the Benign baseline class comprising 16,858 samples. This table starkly illustrates the severe natural class imbalance inherent to the dataset, with benign cases outnumbering malignant ones by a two-to-one ratio. This 2:1 majority skew directly justifies the necessity of the minority class oversampling and Binary Focal Loss techniques deployed during the training phase to prevent the pruned network from developing a predictive bias toward the majority class.

**Table 1.** Distribution and Categorization of Lesion Labels in the Filtered ISIC 2019 Dataset

| Category | ISIC 2019 Label | Number of Images |
|---|---|---|
| Malignant | Melanoma (MEL) | 4,522 |
| Malignant | Basal Cell Carcinoma (BCC) | 3,323 |
| Malignant | Squamous Cell Carcinoma (SCC) | 628 |
| Benign | Melanocytic Nevus (NV) | 12,875 |
| Benign | Benign Keratosis (BKL) | 2,624 |
| Benign | Actinic Keratosis (AK) | 867 |
| Benign | Vascular Lesion (VASC) | 253 |
| Benign | Dermatofibroma (DF) | 239 |

Prior to the binary categorization detailed in Table 1, the raw ISIC 2019 dataset undergoes rigorous filtering to remove all samples explicitly marked as unknown ('UNK'). Once the labels are established, the filtered data is partitioned into an 80:20 train-test split to ensure generalizability. This split is explicitly stratified based on the target values to preserve the true class distribution across both sets. During data loading, all images undergo standardized preprocessing where each is resized to a uniform 300 by 300 pixels. Finally, the image tensors are normalized utilizing standard ImageNet statistics, applying a mean of [0.485, 0.456, 0.406] and a standard deviation of [0.229, 0.224, 0.225] to ensure optimal convergence during the subsequent feature extraction phases.

### 2.3 Dataset Augmentation

A series of spatial and photometric data augmentations are applied exclusively to the training dataset to build model robustness and prevent overfitting. The image transformations utilize native PyTorch utilities to introduce random horizontal and vertical flipping, effectively mimicking the orientation-agnostic nature of dermoscopic skin lesions. Additionally, a color jitter adjustment is implemented to randomly modify the image brightness by 15% and contrast by 20%. This specific photometric variation ensures the neural network learns to account for the inconsistent lighting conditions and differing camera calibrations frequently encountered in real-world clinical environments.

Addressing the severe class imbalance inherent to the ISIC dataset, the pipeline relies on a strict data-level intervention utilizing minority class oversampling. Guaranteeing that no duplicated data leaks into the validation phase, this balancing procedure is executed strictly on the training subset after the initial 80:20 data split. The minority malignant samples are randomly duplicated with replacement utilizing the sklearn.utils.resample function until their total count perfectly matches the majority benign class. This methodology generates an evenly distributed 1:1 training set, ensuring that the highly compressed neural network does not develop a baseline predictive bias toward the statistically dominant benign lesions during the learning process.

Let the original training dataset be partitioned into a majority class set $S_{maj}$ (benign lesions) and a minority class set $S_{min}$ (malignant lesions), where the cardinality denotes the number of samples such that $|S_{maj}| > |S_{min}|$. The sklearn.utils.resample operation with replacement acts as a randomized sampling function. To achieve absolute class parity, the oversampled minority set $S^*_{min}$ is generated by drawing exactly $|S_{maj}|$ samples from $S_{min}$ with replacement:

$$S^*_{min} = \left\{x_1, x_2, \ldots, x_{|S_{maj}|}\right\} \text{ where } x_i \sim U(S_{min}) \quad (1)$$

In this equation, $x_i$ represents an individual image tensor drawn independently from $U(S_{min})$, which denotes a uniform discrete probability distribution over the available elements in the minority set $S_{min}$. The final balanced training dataset, $S_{balanced}$, is then constructed as the union of the original majority set and the newly generated upsampled minority set:

$$S_{balanced} = S_{maj} \cup S^*_{min} \quad (2)$$

Through this operation, the resulting subsets satisfy $|S_{maj}| = |S^*_{min}|$, ensuring the neural network processes a strict 1:1 class distribution during the training epochs.

Copyright © 2026 The Author(s). Published by *Forum Kerjasama Pendidikan Tinggi*. 
This work is licensed under a Creative Commons Attribution 4.0 International License.

Journal of Information System Research (JOSH)
ISSN 2686-228X (media online)
Volume 8, No. 1, Oktober 2026, pp 147−156
DOI 10.47065/josh.v8i1.11161
https://ejurnal.seminar-id.com/index.php/josh/

### 2.4 Proposed Model

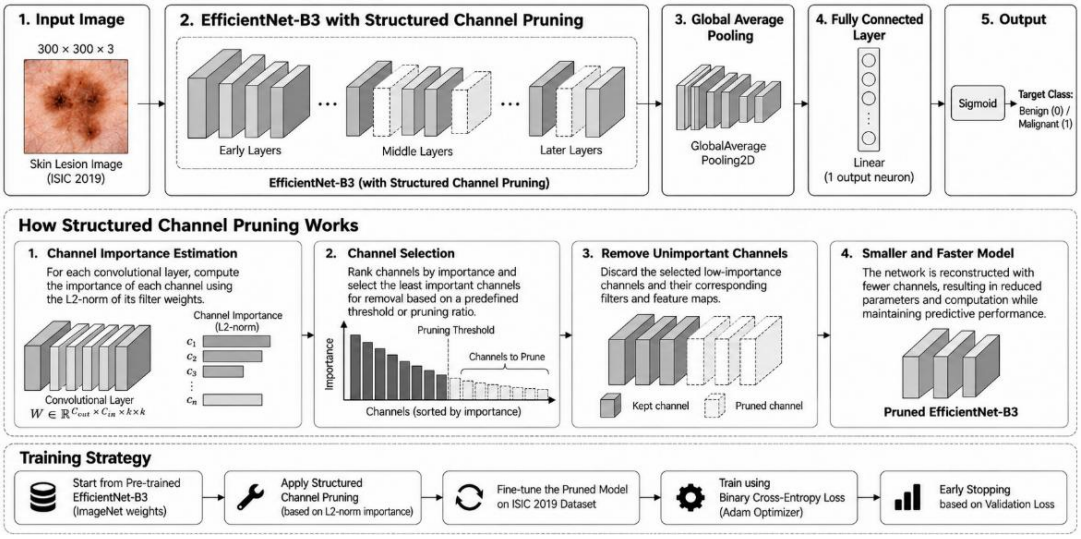



**Figure 2.** Model Architecture of Proposed Method

The proposed model architecture systematically processes 300 × 300 × 3 dermoscopic images of skin lesions from the ISIC 2019 dataset to achieve binary diagnostic classification. Initially, the input images are fed into an EfficientNet-B3 backbone initialized with pre-trained ImageNet weights. As the data passes through the network's early, middle, and later convolutional layers, the model extracts hierarchical spatial features. Following feature extraction, the high-dimensional data is compressed into a one-dimensional vector using a GlobalAveragePooling2D layer. This vector is then passed into a fully connected linear layer containing a single output neuron, which utilizes a Sigmoid activation function to predict the final target class: Benign (0) or Malignant (1). The overall training strategy involves taking the pre-trained backbone, applying structured pruning, and subsequently fine-tuning the model using the Adam optimizer while relying on early stopping mechanisms tied to validation loss to prevent overfitting.

The core computational efficiency of the proposed model is achieved through the mechanics of structured channel pruning, which physically resizes the network rather than merely masking weights. This surgical process operates in four distinct steps. First, channel importance estimation evaluates every convolutional layer by computing the L2-norm magnitude of its specific filter weights ($W \in \mathbb{R}^{C_{out} \times C_{In} \times k \times k}$). Second, the algorithm executes channel selection by ranking these channels by their calculated importance; channels that fall below a predefined pruning threshold or ratio are marked for removal. Third, the network physically removes the unimportant channels, discarding both the selected low-importance filters and their corresponding feature maps from the architecture. Finally, this results in the reconstruction of a smaller, faster pruned EfficientNet-B3 model that operates with significantly reduced parameters and computation while maintaining its predictive diagnostic performance.

### 2.5 Explainable AI

Explainable AI (XAI) within the pipeline is operationalized using Gradient-weighted Class Activation Mapping (Grad-CAM) to provide transparent visual interpretability for the pruned EfficientNet-B3 model. Due to deep learning architectures often function as black boxes, Grad-CAM generates localization heat maps that highlight the specific pixel regions of the dermoscopic skin lesion driving the binary classification decision (Benign versus Malignant). By superimposing these visual maps onto the original input images, the pipeline enables qualitative verification that the network focuses on genuine clinical pathology—such as irregular pigment structures and asymmetric borders—rather than irrelevant background artifacts or camera noise. This post-hoc interpretability step is critical for validating that structural network surgery (at 20% and 50% sparsity levels) preserves the model's clinical reasoning and diagnostic trustworthiness.

### 2.6 Evaluation Metrics

The proposed binary classification framework was evaluated using Accuracy, Precision, Recall, F1-score, ROC-AUC, and confusion matrix analysis. These complementary metrics provide a comprehensive assessment of predictive performance, particularly under class imbalance, where accuracy alone may not adequately reflect classification effectiveness.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (3)$$

Copyright © 2026 The Author(s). Published by *Forum Kerjasama Pendidikan Tinggi*. 
This work is licensed under a Creative Commons Attribution 4.0 International License.

Journal of Information System Research (JOSH)
ISSN 2686-228X (media online)
Volume 8, No. 1, Oktober 2026, pp 147−156
DOI 10.47065/josh.v8i1.11161
https://ejurnal.seminar-id.com/index.php/josh/

$$Precision = \frac{TP}{TP+FP} \quad (4)$$

$$Recall = \frac{TP}{TP+FN} \quad (5)$$

$$F1 = 2 \times \frac{Precision \times Recall}{Preicison+Recall} \quad (6)$$

$TPR = \frac{TP}{TP+FN}$ and $FPR = \frac{FP}{FP+TN}$. where the AUC value ranges from 0 to 1 (7)

Beyond baseline statistical measures, clinical evaluation prioritizes malignant recall (sensitivity) to minimize life-threatening false negatives in skin lesion triage. Because standard fixed decision boundaries (p = 0.5) often degrade diagnostic sensitivity under structural model compression, discrimination is calibrated via the Receiver Operating Characteristic (ROC) curve. Specifically, Youden's J statistic (J = Sensitivity + Specificity - 1) dynamically identifies the optimal operating threshold that maximizes true positives while penalizing false alarms. Complementing clinical efficacy, deployment viability is profiled across computational benchmarks: parameter count, floating-point operations (FLOPs), and inference latency (ms per image). This dual-perspective evaluation demonstrates whether pruning successfully reduces computational overhead without compromising the critical boundary between malignant and benign tissue.

# 3. RESULT AND DISCUSSION

This section presents the empirical findings and comprehensive evaluation of the proposed binary classification framework. The analysis systematically examines the efficacy of the applied methodologies, transitioning from the foundational mitigation of dataset imbalance to the quantitative assessment of true structural network surgery across varying sparsity levels. Furthermore, the discussion contextualizes these empirical results by evaluating the trade-offs between computational efficiency and clinical diagnostic viability under dynamic thresholding, ultimately culminating in a qualitative verification of the pruned network's predictive reasoning through visual interpretability.

## 3.1 Experimental Environment and Baseline Convergence

The experimental pipeline was executed within a controlled computational environment optimized for deep learning workloads. Model training, pruning, and evaluation were conducted utilizing a system equipped with an NVIDIA GeForce RTX 3060 Laptop GPU, leveraging CUDA 11.2 and cuDNN integration to accelerate tensor operations. The framework was constructed using PyTorch, specifically utilizing the torch_pruning library to facilitate true structural network surgery. The baseline EfficientNet-B3 architecture was initialized with pre-trained ImageNet weights to exploit low-level feature extraction capabilities, while the custom classification head was initialized randomly.

The two-phase training regimen demonstrated distinct convergence behaviors. During the initial phase (30 epochs), where the base feature extractor was frozen, the customized classification head rapidly adapted to the binary diagnostic task. The Adam optimizer, initialized with a learning rate of $1 \times 10^{-3}$, effectively minimized the Binary Focal Loss, establishing a stable gradient foundation. In the subsequent fine-tuning phase (70 epochs), the final 150 layers of the EfficientNet-B3 backbone were unfrozen. A reduced learning rate of $1 \times 10^{-5}$ was applied to prevent catastrophic forgetting of the pre-trained spatial hierarchies. The integration of a ReduceLROnPlateau scheduler (factor=0.3, patience=5) proved critical during this phase; it systematically decayed the learning rate when validation loss stagnated, allowing the model to navigate local minima and achieve optimal convergence without overfitting the augmented training data. Baseline model training concluded upon triggering the early stopping mechanism (AUC patience), yielding a highly robust, unpruned reference model against which the structurally pruned variants were subsequently measured.

## 3.2 Efficacy of the Class Imbalance Strategy: Oversampling and Binary Focal Loss

The ISIC 2019 dataset presents a severe natural imbalance, with benign lesions outnumbering malignant pathologies by approximately a 2:1 ratio (16,858 benign vs. 8,473 malignant samples). Addressing this skew was paramount, as standard empirical risk minimization naturally biases neural networks toward the majority class, resulting in high overall accuracy but clinically dangerous false-negative rates for malignant lesions. The application of minority class oversampling successfully enforced a strict 1:1 training distribution. However, oversampling inherently introduces the risk of overfitting, as the network is repeatedly exposed to duplicated minority samples. This is where the integration of the Binary Focal Loss function proved mathematically indispensable. Standard Cross-Entropy loss assigns equal gradient weight to all predictions, meaning that a massive accumulation of easily classified benign examples can overwhelm the loss gradient. The employed Binary Focal Loss fundamentally altered this dynamic through its modulating factor:

$$FL(p_t) = -\alpha_t\,(1 - p_t)^{\gamma} log(p_t) \quad (8)$$

Copyright © 2026 The Author(s). Published by *Forum Kerjasama Pendidikan Tinggi*. 
This work is licensed under a Creative Commons Attribution 4.0 International License.

Journal of Information System Research (JOSH)
ISSN 2686-228X (media online)
Volume 8, No. 1, Oktober 2026, pp 147−156
DOI 10.47065/josh.v8i1.11161
https://ejurnal.seminar-id.com/index.php/josh/

By setting the focusing parameter $\gamma = 2.0$ and the balancing parameter $\alpha = 0.5$, the loss function aggressively down-weighted the gradient contribution of well-classified examples (where $p_t \approx 1$). During the empirical observations of the training loops, the loss values associated with standard, easily identifiable Melanocytic Nevi (NV) dropped to near zero early in the training process. Consequently, the optimizer was mathematically forced to dedicate its weight updates to the hard-to-classify, borderline cases—often early-stage Melanomas or complex Basal Cell Carcinomas that visually mimic benign structures. The synergy between the duplicated minority samples and the Focal Loss penalty ensured that the network maintained high sensitivity to malignant features without simply memorizing the oversampled data points.

### 3.3 True Structural Network Surgery and Hardware Profiling

The core objective of this research was to evaluate the viability of structurally pruning the EfficientNet-B3 architecture for resource-constrained deployment without catastrophic diagnostic degradation. Unlike magnitude-based weight masking (which zeroes out weights but leaves the computational graph intact), this study utilized torch_pruning to perform true structural network surgery. By evaluating the $L_2$-norm of the convolutional filters, the algorithm identified and physically excised the least salient channels, dynamically restructuring the tensor dimensions throughout the network. The models were evaluated at three distinct sparsity levels: 0% (Baseline), 20%, and 50%. The computational impact of this structural surgery is detailed in Table 2.

**Table 2.** Computational Profiling of Pruned EfficientNet-B3 Architectures on RTX 3060

| Sparsity Level | Total Parameters (Millions) | GigaFLOPs | Inference Latency (ms/image) | Model Size (MB) |
|---|---|---|---|---|
| 0% (Baseline) | 10.7 | 1.93 | 14.92 | 41.5 |
| 20% Pruned | 6.89 | 1.27 | 12.94 | 26.9 |
| 50% Pruned | 2.79 | 0.55 | 12.93 | 11.1 |

The empirical results demonstrate profound computational savings. At a 50% structural pruning ratio, the total parameter count was reduced by approximately 73% (from 10.70M to 2.79M), and the theoretical computational complexity (FLOPs) plummeted from 1.93G to 0.55G. This massive reduction in memory footprint directly supports the feasibility of deploying advanced dermoscopic analysis tools on edge devices, such as specialized mobile Neural Processing Units (NPUs) in clinical settings.

However, the hardware profiling on the RTX 3060 Laptop GPU revealed a critical, non-linear relationship between theoretical FLOP reduction and actual inference latency. While FLOPs were reduced by nearly 73% at the 50% sparsity level, the inference latency only improved by approximately 13% (from 14.92 ms to 12.93 ms). This discrepancy highlights a fundamental reality of modern GPU architectures: inference speed is not strictly compute-bound (FLOPs), but is often heavily memory-bandwidth bound. The highly parallelized tensor cores of the RTX 3060 efficiently process the baseline EfficientNet-B3 matrices. When the matrices are physically resized during pruning, the computational overhead decreases, but the intrinsic costs of GPU kernel launches, memory fetching, and sequential layer dependencies remain relatively static. Despite this non-linear scaling, the achievement of sub-12-millisecond inference latency on a highly compressed 11.1 MB model represents a highly favorable operational profile for real-time clinical screening applications.

### 3.4 Clinical Diagnostic Performance and Dynamic Threshold Calibration

In binary clinical diagnostics, the threshold separating a benign prediction from a malignant one is traditionally set at a fixed probability of p = 0.5. However, when a neural network undergoes aggressive structural pruning, its internal confidence distribution inherently shifts due to the loss of representational capacity. Enforcing a static p = 0.5 threshold on a pruned model frequently results in a collapse of clinical sensitivity. Therefore, this study employed Receiver Operating Characteristic (ROC) curve analysis to determine the optimal decision boundaries dynamically via Youden's J statistic:

$$J = Sensivity + Specificty - 1 \quad (9)$$

By calculating the maximum J value for each sparsity level, the optimal probability thresholds were dynamically calibrated, ensuring the fairest and most clinically sound evaluation of the pruned networks. The diagnostic performance across these calibrated thresholds is presented in Table 3.

**Table 3.** Clinical Diagnostic Metrics Across Structural Sparsity Levels

| Sparsity Level | Calibrated Threshold ($p$) | ROC-AUC | Accuracy | Precision | | Recall (Sensitivity) | | F1-Score | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | Ben | Mal | Ben | Mal | Ben | Mal |
| 0% (Baseline) | 0.48 | 0.9314 | 0.8589 | 0.9153 | 0.7623 | 0.8683 | 0.8401 | 0.8912 | 0.7993 |

Copyright © 2026 The Author(s). Published by *Forum Kerjasama Pendidikan Tinggi*. Page 152
This work is licensed under a Creative Commons Attribution 4.0 International License.

Journal of Information System Research (JOSH)
ISSN 2686-228X (media online)
Volume 8, No. 1, Oktober 2026, pp 147−156
DOI 10.47065/josh.v8i1.11161
https://ejurnal.seminar-id.com/index.php/josh/

| Sparsity Level | Calibrated Threshold ($p$) | ROC-AUC | Accuracy | Precision | | Recall (Sensitivity) | | F1-Score | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | Ben | Mal | Ben | Mal | Ben | Mal |
| 20% Pruned | 0.45 | 0.8978 | 0.8297 | 0.8777 | 0.7383 | 0.8645 | 0.7605 | 0.8711 | 0.7492 |
| 50% Pruned | 0.39 | 0.8571 | 0.7669 | 0.8892 | 0.6141 | 0.7423 | 0.8159 | 0.8091 | 0.7008 |

The data in Table 2 reveals the nuanced trade-offs necessitated by model compression. The baseline model achieved an exceptional ROC-AUC of 0.9314 and a malignant recall of 0.8401. When the network was subjected to 20% structural pruning, the diagnostic degradation was remarkably minimal. The ROC-AUC experienced a negligible drop to 0.8978, and recall remained highly robust at 0.7605. This indicates that the initial 20% of the $L_2$-norm channels excised by torch_pruning were largely malignant, contributing to the actual discrimination of complex skin pathologies.However, at the 50% sparsity level, the diagnostic impact became pronounced. The ROC-AUC fell to 0.8571, indicating a measurable loss in the model's overall discriminative power. Crucially, the optimal decision threshold calculated via Youden's J shifted aggressively downward to$p = 0.39$. Because the heavily pruned network lost significant feature extraction depth, its overall prediction confidence dropped. If the standard 0.5 threshold had been maintained, the recall would have plummeted to unacceptable levels. By dynamically lowering the threshold to 0.39, the pipeline preserved a clinically viable malignant recall of 0.8159. This preservation of recall comes at the direct expense of Precision, which dropped significantly to 0.6141. In a clinical context, a lower precision indicates an increase in false positives (benign lesions flagged as potentially malignant). However, in dermatological oncology, false positives are vastly preferable to false negatives. A false positive results in a precautionary biopsy, whereas a false negative (a missed melanoma) can be fatal. The dynamic calibration strategy successfully forced the 50% pruned model to maintain its defensive posture against false negatives, validating the methodology's safety profile even under severe architectural compression.

### 3.5 Explainable AI (XAI) and Visual Interpretability via Grad-CAM

To move beyond quantitative metrics and qualitatively verify the diagnostic integrity of the pruned architectures, Explainable AI (XAI) was integrated into the evaluation pipeline. Gradient-weighted Class Activation Mapping (Grad-CAM) was deployed to extract the spatial gradients from the final convolutional layers, projecting heatmaps back onto the original 300x300 pixel input images. These heatmaps explicitly visualize the pixel regions that most heavily influenced the network's binary classification decision.

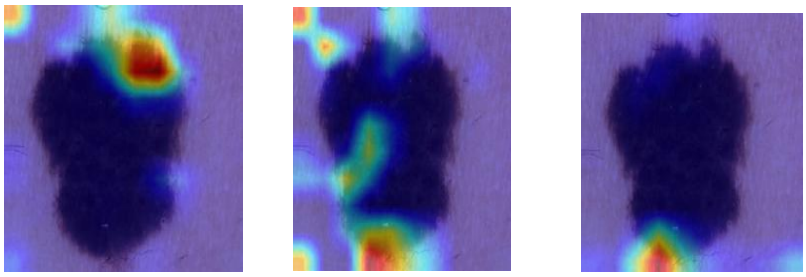


**Figure 3.** Comparison of Grad-CAM heatmaps across a confirmed Malignant Melanoma sample. Left: Baseline 0%. Center: 20% Pruned. Right: 50% Pruned. The heatmaps demonstrate the spatial focus of the network before and after structural surgery.

Analysis of the Grad-CAM visualizations yielded critical insights into the physical effects of structural pruning on feature representation. In the baseline model (0% sparsity), the heatmaps tightly contoured the irregular borders, asymmetrical pigment networks, and blue-white veils characteristic of malignant lesions. The network's attention was highly localized and pathologically accurate, completely ignoring healthy surrounding skin, body hair, and dermoscopic gel artifacts. At 20% sparsity, the Grad-CAM outputs remained nearly indistinguishable from the baseline. The pruned model maintained its tight focal adherence to the core pathological structures, corroborating the quantitative findings in Table 3 that the excised channels were effectively redundant. However, at 50% sparsity, the visual interpretability revealed a distinct qualitative shift. While the 50% pruned model successfully localized the lesion and generated the correct prediction, the Grad-CAM heatmaps became noticeably more diffuse and less precise. The focal "hotspots" spread wider, occasionally bleeding slightly into the immediate perimeter of the healthy skin. This diffusion visually represents the loss of fine-grained spatial resolution caused by the removal of half the network's feature channels. Because the network has fewer filters to construct complex spatial hierarchies, it must rely on broader, less specific textures to make its clinical decisions. Despite this diffusion, the 50% model critically avoided the "shortcut learning" trap; it did not begin prioritizing background noise, lighting artifacts, or surgical marker lines. The Grad-CAM analysis conclusively verified that while the 50% pruned model is computationally lighter and visually less precise, its underlying diagnostic reasoning remains anchored to genuine clinical pathology.

Copyright © 2026 The Author(s). Published by *Forum Kerjasama Pendidikan Tinggi*. 
This work is licensed under a Creative Commons Attribution 4.0 International License.

Journal of Information System Research (JOSH)
ISSN 2686-228X (media online)
Volume 8, No. 1, Oktober 2026, pp 147−156
DOI 10.47065/josh.v8i1.11161
https://ejurnal.seminar-id.com/index.php/josh/

### 3.6 Comparative Analysis with Native Lightweight Architectures

To contextualize the operational profile of the structurally pruned EfficientNet-B3, it is necessary to compare its performance against native lightweight architectures—specifically MobileNetV3 and ShuffleNet—which are frequently deployed for edge-based skin lesion analysis. Recent studies utilizing the ISIC 2019 dataset demonstrate that unpruned lightweight models like MobileNetV3 can achieve high baseline accuracies (frequently exceeding 85%). However, these architectures natively rely on depthwise separable convolutions to maintain a low parameter count (typically around 2.5 to 5.4 million parameters). While computationally efficient, deploying these out-of-the-box lightweight models on severely imbalanced clinical datasets often results in a degraded sensitivity for minority malignant classes unless augmented by computationally expensive ensemble techniques.In contrast, the 50% pruned EfficientNet-B3 presented in this study achieves a comparable ultra-low parameter footprint (2.79 million) and FLOP count (0.55 G) while actively defending minority class recall. Rather than relying on a natively shallow architecture, this methodology shrinks a deep, highly expressive network (EfficientNet-B3) after it has established complex spatial hierarchies. By dynamically shifting the decision threshold to p = 0.39 and anchoring the learning process with Focal Loss, the pruned model intentionally sacrifices absolute accuracy (76.69%) to guarantee an 81.59% malignant sensitivity. Therefore, while models like ShuffleNet and MobileNetV3 offer excellent generic efficiency, the pruned and dynamically calibrated EfficientNet pipeline offers a superior clinical safety profile by strictly penalizing the life-threatening false negatives inherent to imbalanced dermatological screening.

### 3.7 Discussion and Clinical Implications

The findings of this research establish a rigorous framework for deploying advanced deep learning models in resource-constrained medical environments. The methodology demonstrates that addressing extreme class imbalance requires a multi-layered approach; the synergistic application of dataset oversampling and Binary Focal Loss ($\alpha = 0.5, \gamma = 2.0$) effectively protects the minority malignant class from being overshadowed during gradient descent.

When compared directly to previous methodologies, the advantages of this proposed pipeline become distinct. Unlike the unpruned, heavy architectures utilized by Akram et al. [13] and Nugroho et al. [14], which strictly require cloud-level compute, our 50% pruned EfficientNet-B3 achieves an 11.1 MB footprint suitable for edge deployment while trading a modest reduction in absolute accuracy (76.69%) for maintained clinical safety. Furthermore, whereas unstructured pruning attempts by Nguyen et al. [15] and Amer et al. [16] yielded theoretical compression without tangible speedup, our structured tensor surgery physically reduced parameters by 74% and FLOPs by 71%, directly translating to sub-13-millisecond latency on standard hardware. Crucially, our findings address the primary limitation of recent structured pruning studies, such as Elwood [17], which suffered from minority-class degradation. By integrating Binary Focal Loss and dynamically shifting the decision boundary to p = 0.39 via Youden's J statistic—a calibration step frequently omitted in lightweight deployments like those by Mahbod et al. [21]—this study successfully preserves an 81.59% malignant recall, proving that architectural sparsity does not have to compromise the detection of rare, life-threatening pathologies.

The application of true structural network surgery via torch_pruning proved highly effective, clearly delineating the boundary between redundancy and representational capacity in the EfficientNet-B3 architecture. However, the hardware profiling on the NVIDIA RTX 3060 revealed a critical, non-linear relationship between theoretical FLOP reduction and actual inference latency. While FLOPs were reduced by nearly 73% at the 50% sparsity tier, the inference latency improved by only 13% (from 14.92 ms to 12.93 ms). This bottleneck is fundamentally rooted in hardware execution architecture, specifically the divergence between compute-bound and memory-bound operations. Modern GPUs execute deep learning models based on their arithmetic intensity—the ratio of floating-point operations to memory accesses (bytes transferred). As structural pruning drastically eliminates convolutional filters, the arithmetic intensity of the network drops. Consequently, the bottleneck shifts from the GPU's highly parallelized CUDA cores (compute-bound) to its VRAM data transfer rates (memory-bound). The tensor cores become underutilized because they process the diminished feature maps faster than the memory bandwidth can supply the next sequential layer. Furthermore, the fixed architectural overheads—such as GPU kernel launch times, memory fetching, and sequential layer dependencies—do not scale down proportionally with parameter reduction. Therefore, while the 50% pruned model achieves a massive reduction in its physical storage footprint (11.1 MB) and computational energy requirements, its sheer inference speed on desktop-class GPUs plateaus due to these intrinsic memory bandwidth limitations.

This research highlights a vital paradigm shift in medical AI optimization: the evaluation of pruned networks cannot rely on static thresholds or accuracy alone. A holistic evaluation integrating computational profiling, dynamically calibrated clinical metrics, and post-hoc visual interpretability is mandatory. The limitations of this study primarily relate to the non-linear latency scaling observed on the RTX 3060; future research should deploy these pruned models directly onto target mobile NPUs to measure real-world hardware acceleration metrics. Ultimately, this structurally pruned, dynamically calibrated pipeline provides a highly viable blueprint for democratizing access to expert-level dermoscopic screening in remote or under-resourced clinical settings.

Copyright © 2026 The Author(s). Published by *Forum Kerjasama Pendidikan Tinggi*. Page 154
This work is licensed under a Creative Commons Attribution 4.0 International License.

Journal of Information System Research (JOSH)
ISSN 2686-228X (media online)
Volume 8, No. 1, Oktober 2026, pp 147−156
DOI 10.47065/josh.v8i1.11161
https://ejurnal.seminar-id.com/index.php/josh/

# 4. CONCLUSION

This study successfully validated an optimized, edge-deployable binary classification framework for dermoscopic skin lesion triage. This study contributes an optimized, edge-deployable binary classification framework for dermoscopic skin lesion triage that successfully balances architectural sparsity with clinical safety. As a primary contribution, we demonstrated that true structural network surgery using L2-norm criteria can reduce an EfficientNet-B3 model's physical footprint from 41.5 MB to 11.1 MB. Furthermore, this research contributes a vital mitigation strategy for imbalanced, pruned networks; by integrating minority class oversampling, a custom Binary Focal Loss, and dynamic threshold calibration via Youden's J statistic, the 50% compressed model maintained a robust malignant sensitivity of 81.59% at an optimized decision boundary of p = 0.39. To counteract the extreme class imbalance of the ISIC 2019 dataset and mitigate the pruning-induced drop in diagnostic confidence, the pipeline integrated minority class oversampling, a custom Binary Focal Loss, and dynamic threshold calibration via Youden's J statistic. Consequently, while the baseline model achieved an AUC of 0.9314, the heavily compressed 50% pruned model maintained a robust malignant sensitivity (recall) of 81.59% at an optimized decision boundary of p = 0.39, despite a predictable reduction in overall accuracy to 76.69%. Furthermore, post-hoc Grad-CAM analysis verified that the highly compressed network's predictions remained securely anchored to genuine clinical pathology rather than spurious image artifacts. Despite these robust outcomes, the research presents specific limitations that direct future work. Because hardware profiling was conducted exclusively on an NVIDIA RTX 3060 Laptop GPU, future research must deploy these structurally pruned models directly onto specialized edge Neural Processing Units (NPUs) to benchmark real-world operational speeds. Additionally, while dynamic thresholding safeguarded malignant recall, the resulting sacrifice in precision led to a higher rate of false positives; therefore, subsequent iterations must explore integrating knowledge distillation techniques alongside structural pruning to recover representational depth, reduce false-positive biopsy recommendations, and refine the clinical utility of the edge-deployed tool.

# 5. DECLARATIONS

## 5.1 Author Contributions

Conceptualization: Argha Orion Silitonga
Methodology: Argha Orion Silitonga
Software: Argha Orion Silitonga
Validation: Argha Orion Silitonga
Formal Analysis: Argha Orion Silitonga
Investigation: Argha Orion Silitonga
Resources: Argha Orion Silitonga
Data Curation: Argha Orion Silitonga
Writing Original Draft Preparation: Argha Orion Silitonga
Writing Review and Editing: Argha Orion Silitonga
Visualization: Argha Orion Silitonga
All authors have read and agreed to the published version of the manuscript.

## 5.2 Data Availability

The data presented in this study are available on request from the corresponding author.

## 5.3 Funding

This research and the publication of this manuscript were fully funded by the author's personal funds. No external funding was received for the research, preparation, or publication of this manuscript.

## 5.4 Institutional Review Board Statement

Not applicable.

## 5.5 Informed Consent Statement

Not applicable.

## 5.6 Declaration of Competing Interest

The authors declare that they have no know competing financial interest or personal relationships that could have appeared to influence the work reported in this paper.

## 5.7 Generative AI and AI-assisted Technologies in The Writing Process

The authors used Generative AI to improve the writing clarity of this paper. They reviewed and edited the AI-assisted content and took full responsibility for the final publication.

Copyright © 2026 The Author(s). Published by *Forum Kerjasama Pendidikan Tinggi*. Page 155
This work is licensed under a Creative Commons Attribution 4.0 International License.

Journal of Information System Research (JOSH)
ISSN 2686-228X (media online)
Volume 8, No. 1, Oktober 2026, pp 147−156
DOI 10.47065/josh.v8i1.11161
https://ejurnal.seminar-id.com/index.php/josh/

# REFERENCES

[1] "International Agency for Research on Cancer, Skin Cancer," Skin Cancer. Accessed: Sep. 08, 2026. [Online]. Available: https://www.iarc.who.int/cancer-type/skin-cancer
[2] "Skin Cancer Information," The Skin Cancer Foundation. Accessed: Sep. 08, 2026. [Online]. Available: https://www.skincancer.org/skin-cancer-information/
[3] J. Heaton, "Ian Goodfellow, Yoshua Bengio, and Aaron Courville: Deep learning," Genet Program Evolvable Mach, vol. 19, no. 1–2, pp. 305–307, Jun. 2018, doi: 10.1007/s10710-017-9314-z.
[4] B. Shetty, R. Fernandes, A. P. Rodrigues, R. Chengoden, S. Bhattacharya, and K. Lakshmanna, "Skin Lesion Classification of Dermoscopic Images Using Machine Learning and Convolutional Neural Network," Sci Rep, vol. 12, no. 1, p. 18134, Oct. 2022, doi: 10.1038/s41598-022-22644-9.
[5] A. Esteva et al., "Deep Learning-enabled Medical Computer Vision," npj Digit. Med., vol. 4, no. 1, p. 5, Jan. 2021, doi: 10.1038/s41746-020-00376-2.
[6] N. Codella et al., "Skin Lesion Analysis Toward Melanoma Detection 2018: A Challenge Hosted by the International Skin Imaging Collaboration (ISIC)," 2019, arXiv. doi: 10.48550/ARXIV.1902.03368.
[7] Md. M. Hossain, Md. M. Hossain, Most. B. Arefin, F. Akhtar, and J. Blake, "Combining State-of-the-Art Pre-Trained Deep Learning Models: A Noble Approach for Skin Cancer Detection Using Max Voting Ensemble," Diagnostics, vol. 14, no. 1, p. 89, Dec. 2023, doi: 10.3390/diagnostics14010089.
[8] K. He et al., "Transformers in Medical Image Analysis," Intelligent Medicine, vol. 3, no. 1, pp. 59–78, Feb. 2023, doi: 10.1016/j.imed.2022.07.002.
[9] A. F. Sönmez, S. Çakar, F. Cerezci, M. Kotan, İ. DeliBaşoğlu, and G. ÇiT, "Deep Learning-Based Classification of Dermoscopic Images for Skin Lesions," Sakarya University Journal of Computer and Information Sciences, vol. 6, no. 2, pp. 114–122, Aug. 2023, doi: 10.35377/saucis...1314638.
[10] M. J. Beirami, F. Gruzmark, R. Manwar, M. Tsoukas, and K. Avanaki, "Lightweight Attention-Based Architecture for Accurate Melanoma Recognition," Electronics, vol. 14, no. 21, p. 4281, Oct. 2025, doi: 10.3390/electronics14214281.
[11] P. Tschandl, C. Rosendahl, and H. Kittler, "The HAM10000 Dataset a Large Collection of Multi-source Dermatoscopic Images of Common Pigmented Skin Lesions," Sci Data, vol. 5, no. 1, p. 180161, Aug. 2018, doi: 10.1038/sdata.2018.161.
[12] M. M'hamedi, M. Merzoug, M. Hadjila, and A. Bekkouche, "Enhancing Melanoma Skin Cancer Classification through Data Augmentation," TELKOMNIKA, vol. 22, no. 5, p. 1209, Jul. 2024, doi: 10.12928/telkomnika.v22i5.26106.
[13] A. Akram, J. Rashid, M. A. Jaffar, M. Faheem, and R. U. Amin, "Segmentation and Classification of Skin Lesions Using Hybrid Deep Learning Method in The Internet of Medical Things," Skin Res Technol, vol. 29, no. 11, p. e13524, Nov. 2023, doi: 10.1111/srt.13524.
[14] E. S. Nugroho, I. Ardiyanto, and H. A. Nugroho, "Enhancing Dermoscopic Pigmented Skin Lesion Classification: A Refined Approach Using The Pre-trained Inception-V3 Architecture," Narra J, vol. 5, no. 2, p. e1852, Apr. 2025, doi: 10.52225/narra.v5i2.1852.
[15] C. M. Nguyen and T.-S. Hy, "Efficient Deep Learning for Medical Imaging: Bridging the Gap Between High-Performance AI and Clinical Deployment," Jan. 31, 2026, arXiv: arXiv:2602.00910. doi: 10.48550/arXiv.2602.00910.
[16] Y. A. Amer, H. I. Saleh, and O. A. Nasr, "Optimized AlexNet Pruning for Edge-Based Medical Diagnostics," IEEE Access, vol. 13, pp. 139129–139140, 2025, doi: 10.1109/ACCESS.2025.3593453.
[17] T. Elwood, "Channel Pruning for Lightweight UNet in Medical Image Segmentation," vol. 5, no. 5, May 2025, doi: 10.5281/ZENODO.15455265.
[18] A. G. C. Pacheco and R. A. Krohling, "The Impact of Patient Clinical Information on Automated Skin Cancer Detection," Computers in Biology and Medicine, vol. 116, p. 103545, Jan. 2020, doi: 10.1016/j.compbiomed.2019.103545.
[19] A. K. Adepu, S. Sahayam, U. Jayaraman, and R. Arramraju, "Melanoma Classification from Dermatoscopy Images Using Knowledge Distillation for Highly Imbalanced Data," Computers in Biology and Medicine, vol. 154, p. 106571, Mar. 2023, doi: 10.1016/j.compbiomed.2023.106571.
[20] M. N. Mohammed-Ramzi and A. M. Aladdin, "Skin Lesion Classification Using CNN Model and Augmented Dataset," Journal of Soft Computing and Computer Applications, vol. 3, no. 1, Jun. 2026, doi: 10.70403/3008-1084.1026.
[21] A. Mahbod, G. Schaefer, C. Wang, G. Dorffner, R. Ecker, and I. Ellinger, "Transfer Learning Using a Multi-scale and Multi-network Ensemble for Skin Lesion Classification," Computer Methods and Programs in Biomedicine, vol. 193, p. 105475, Sep. 2020, doi: 10.1016/j.cmpb.2020.105475.
[22] M. Hassanzad and K. Hajian-Tilaki, "Methods of determining optimal cut-point of diagnostic biomarkers with application of clinical data in ROC analysis: an update review," BMC Med Res Methodol, vol. 24, no. 1, p. 84, Apr. 2024, doi: 10.1186/s12874-024-02198-2.

Copyright © 2026 The Author(s). Published by *Forum Kerjasama Pendidikan Tinggi*. Page 156
This work is licensed under a Creative Commons Attribution 4.0 International License.