The International Society of Data Scientists
Journal of Data Science and Artificial Intelligence
Vol. 5, No. 1. Online ISSN: 2831-4794
Available at https://www.isods.org/publication/

Skin Cancer Detection on Smartphone Images through Knowledge Distillation of Multimodal Deep Learning Models

Nhan Le Thanh Pham[a,b], Duc Dat Pham[*a,b]
[a]International University, Ho Chi Minh City, Vietnam; [b]Vietnam National University, Ho Chi Minh City, 700000, Vietnam;
*Corresponding author

ABSTRACT

Skin cancer remains a critical global health challenge, particularly in regions with limited access to specialized diagnostic tools. This study presents an innovative approach for skin lesion classification using non-dermoscopic smartphone images, leveraging knowledge distillation to enhance model efficiency and accuracy. We utilize the PAD-UFES-20 dataset, which comprises 2,298 smartphone-captured images of six distinct skin lesion types, accompanied by comprehensive patient metadata. Our methodology involves a teacher–student framework, where a ConvNeXt-based teacher model integrated with Convolutional Block Attention Modules (CBAM) and a metadata encoder transfers its learned representations to a more compact EfficientNet-B0 student model. The distillation process incorporates logit matching, feature similarity, and attention transfer, enabling the student model to achieve performance parity with the teacher while significantly reducing computational overhead. Experimental results demonstrate that the student model attains an accuracy of 80.43% and a weighted F1-score of 80.16%, closely mirroring the teacher's performance. Additionally, the integration of metadata and attention mechanisms substantially improves classification robustness, particularly for underrepresented lesion categories. The proposed framework effectively addresses class imbalance through the application of focal loss, enhancing the model's ability to detect clinically significant but less frequent skin lesions. This approach offers a viable solution for deploying accurate skin cancer diagnostic tools on resource-constrained mobile devices, thereby expanding access to essential healthcare services in underserved communities.

Keywords: Skin Cancer Detection, Smartphone Imaging, Knowledge Distillation, Deep Learning, Multimodal Models, Convolutional Neural Networks, Metadata Integration, Mobile Health

1. INTRODUCTION

Skin cancer continues to be a major global health issue, impacting diverse populations across different demographics and geographic areas [1]. Timely identification and precise categorization of worrisome lesions are crucial for effective therapy and enhanced patient survival rates [2]. Nevertheless, in numerous remote and under-resourced areas, access to specialized diagnostic instruments like dermoscopes or to experienced dermatologists is frequently deficient [3]. This gap highlights the necessity for diagnostic techniques utilizing more accessible equipment, such a smartphone camera, to take photos of skin lesions in less controlled environments.

Recent advancements in deep learning have showcased the capability of convolutional neural networks (CNNs) to accurately detect and categorize skin lesions, particularly when trained on extensive datasets of medical pictures [4]. However, a significant difficulty endures: numerous advanced models continue to be computationally demanding and memory-intensive, creating obstacles for their implementation on resource-constrained devices [5]. This limitation is especially pronounced in rural or low-income regions where internet bandwidth and hardware capabilities may be restricted. Furthermore, conventional modeling methodologies sometimes neglect to incorporate patient-specific attributes, such age, lifestyle factors, and medical history, which could improve model efficacy when integrated with image-based features [6].

This study introduces an innovative knowledge distillation architecture [7] aimed at enhancing the efficiency and accuracy of skin lesion classification using non-dermoscopic smartphone photos. The principal dataset included in this research is PAD-UFES-20 [8], comprising pictures associated with six unique kinds of skin lesions. We intend to maintain predictive performance and reduce computational overhead by utilizing knowledge distillation from a bigger "teacher" model to a smaller "student" model, hence enhancing practicality for on-device execution in resource-constrained environments. Our teacher network utilizes a ConvNeXt architecture [9] enhanced with Convolutional Block

Attention Modules (CBAM) [10] to emphasize significant picture regions, along with a specialized encoder for patient metadata. The student model, utilizing the EfficientNet-B0 backbone [11], acquires essential feature representations via several distillation mechanisms: logit matching, feature similarity, and attention transfer. A targeted loss is utilized to address the prevalent imbalance in lesion classifications and to prioritize clinically important yet underrepresented categories.

This research presents three primary contributions. We provide a metadata-aware teacher-student pipeline that markedly decreases inference complexity while preserving good classification accuracy across the six skin lesion categories in PAD-UFES-20. Secondly, our findings demonstrate that integrating acquired attention mechanisms with patient metadata enhances classification performance and provides insights into the morphological indicators pertinent to dermatological evaluations. Third, we demonstrate how this strategy might be especially advantageous in contexts lacking specialist equipment, so fulfilling a vital requirement for equitable healthcare access in remote or underprivileged communities.

The subsequent sections of this work are structured as follows. **Section 2**: Related Work examines the current research on deep learning for skin cancer classification, attention mechanisms, metadata application, and knowledge distillation in medical contexts. **Section 3**: Methodology outlines the dataset description, preprocessing procedures, model architectures for both the teacher and student models, the knowledge distillation method, and implementation details. **Section 4**: Experiments and Results delineates the experimental configuration and outcomes, including assessment metrics, comparative analysis of model performances, and ablation investigations. **Section 5**: The discussion interprets the findings, explores their ramifications, acknowledges limits, and proposes directions for further research. Section 6: Conclusion encapsulates the study's contributions, emphasizes its importance, and highlights the potential influence on mobile health applications for skin cancer diagnosis.

2. RELATED WORKS

Deep Learning for Skin Lesion Classification. The advent of deep learning, and particularly convolutional neural networks (CNNs), has significantly advanced skin lesion classification, primarily through architectures such as ResNet, DenseNet, and EfficientNet on datasets like ISIC and HAM10000. Recently, P. Yao et al. [12] demonstrated that moderate-complexity DCNN architectures—combined with tailored augmentation strategies (e.g., RandAugment variants) and specialized loss functions (e.g., Multi-Weighted New Loss)—can handle uneven class distributions and reduce overfitting, thereby achieving accuracy comparable or superior to ensemble models while remaining viable for mobile devices. Similarly, L. Hoang et al. [13] proposed a segmentation-first approach, followed by a lightweight wide-ShuffleNet classifier, enabling real-time deployment on low-power platforms without sacrificing performance. Further investigations by S. Benyahia et al. [14] emphasized the potential of systematically examining pre-trained CNN architectures (e.g., DenseNet201) alongside various classifiers, often exceeding 90% accuracy on large-scale benchmarks such as ISIC 2019 and PH2. Collectively, these studies highlight the adaptability of DCNN-based methods across diverse network depths, preprocessing techniques, and computational settings, reinforcing deep learning's efficacy for accurate and efficient skin lesion diagnosis in resource-limited environments. Despite excelling at extracting hierarchical features from high-resolution dermoscopic images—capturing subtle lesion patterns that aid in cancer detection—these models fundamentally rely on large, high-quality dermoscopic datasets, thus limiting their reach in broader, real-world contexts where such specialized imaging tools may not be readily available.

Attention Mechanisms in Medical Imaging. To further enhance diagnostic precision, attention mechanisms have been integrated into deep neural networks to direct the model's focus toward visually critical regions while diminishing background noise. For instance, S. K. Datta et al. [15] proposed a Soft-Attention module that selectively amplifies significant lesion features, yielding notable improvements in precision and sensitivity on datasets such as HAM10000 and ISIC-2017. In a complementary vein, A. G. C. Pacheco and R. A. Krohling [16] introduced a Metadata Processing Block (MetaBlock) that leverages patient metadata to refine essential image-derived features. This approach not only enhances classification results but also underscores the added benefit of incorporating contextual clinical data into attention-driven pipelines. Collectively, these studies illustrate that attention-based refinement—whether spatial or channel-focused—can substantially boost both classification accuracy and model interpretability by highlighting the most diagnostically relevant lesion areas.

Incorporating Metadata in Diagnostic Models. Beyond visual data, the integration of clinical metadata—such as patient demographics, medical history, and lesion-specific attributes—has been shown to significantly enhance the

performance of diagnostic models. Metadata provides contextual information that complements image-based analysis, enabling models to capture correlations between clinical factors and visual manifestations of skin lesions. For example, C. Dong et al. [17] proposed a comprehensive pipeline wherein a Multi-Scale Holistic Feature Exploration (MSH) module enhances segmentation by extracting perceptual cues across multiple scales, while a Cross-Modality Collaborative Feature Exploration (CMC) module jointly exploits visual and clinical information to dynamically capture interrelated features. By combining these modules and cropping the lesion area based on segmentation masks, their approach reaches state-of-the-art accuracy and AUC on multiple ISIC benchmarks, demonstrating both high diagnostic precision and minimal reliance on extensive labeled data. In a related vein, A. Pundhir et al [18] developed a multimodal framework that fuses patient demographics with image-based features, thereby addressing the common issue of missing or unknown patient data. Validated on the PAD-UFES-20 dataset, this method substantially outperformed baseline CNN architectures, underscoring the robustness and efficacy of metadata integration. This integration not only enriches the feature space but also aligns with clinical practices, where a comprehensive assessment of both visual and contextual factors is essential for accurate diagnosis.

Smartphone-Based and Non-Dermoscopic Imaging Approaches. Recent investigations underscore both the promise and the limitations of using smartphone-acquired images for skin lesion diagnosis. While smartphones offer a low-cost, widely available platform that can facilitate telemedicine and remote screening—potentially reaching patients who lack access to specialized clinical equipment—studies indicate that the diagnostic accuracy may be notably lower when compared to dermoscopic imaging. For instance, A. Dascalu et al. [19] found that smartphone-captured images yielded a sensitivity of around 75.3% for non-melanoma skin cancer (NMSC), falling short of the approximately 95.3% achieved by dermoscopy. Factors such as inconsistent lighting, varying image resolutions, focus issues, and background clutter contribute to diminished model performance.

Knowledge Distillation for Model Deployment. Deploying deep learning models on mobile and edge devices necessitates a balance between computational efficiency and diagnostic accuracy. Knowledge distillation has emerged as a pivotal technique in this regard, enabling the transfer of knowledge from large, high-performing teacher models to smaller, more efficient student models. This process involves training the student model to mimic the output distributions of the teacher model, thereby retaining essential decision-making capabilities while significantly reducing model size and inference time. In medical applications, where both accuracy and efficiency are paramount, knowledge distillation offers a viable pathway to deploy sophisticated diagnostic models on resource-constrained platforms without compromising performance For instance, Y. Wang et al. [20] introduced a “SSD-KD” framework that unifies diverse knowledge—ranging from intra-instance features to softened outputs—into a single training scheme, allowing a distilled MobileNetV2 to attain approximately 85% accuracy on ISIC 2019 while maintaining minimal computing requirements. Similarly, A. K. Adepu et al. [21] combined distillation and cost-sensitive learning with focal loss to produce a lightweight CNN that delivers a 0.8087 sensitivity on ISIC-2020’s melanoma classification task, a marked improvement over non-distilled baselines. In another vein, M. R. Kabir et al. [22] explored multiple distillation strategies from large teacher models (ResNet50 and DenseNet161) to a “TinyStudent” (0.35M parameters), achieving up to 85.45% accuracy yet dramatically reducing memory usage—thereby facilitating deployment on resource-limited medical devices. Lastly, R. Yadav and A. Bhat [23] demonstrated how training a high-accuracy teacher model, then transferring its knowledge to a simpler student network, can preserve a substantial portion of the teacher’s performance (e.g., around 0.80 test accuracy) while shrinking computational demands. Collectively, these studies validate knowledge distillation as an effective method for producing compact yet powerful classification models, suited for real-world skin lesion screening in low-resource or on-device settings.

Research Gaps and Contribution of This Study. While significant progress has been made in applying deep learning to skin lesion classification, most existing research focuses on controlled, high-quality dermoscopic images, limiting the applicability of these models in everyday settings. The integration of attention mechanisms and metadata has shown promise in enhancing model performance, yet their combined application in the context of non-dermoscopic, smartphone-based imaging remains underexplored. Additionally, the deployment of efficient models through knowledge distillation tailored for mobile devices represents a critical advancement towards practical, real-world applications. This study addresses these gaps by leveraging the PAD-UFES-20 dataset—a pioneering collection of smartphone-captured skin lesion images with comprehensive metadata—and employing a teacher–student framework to develop models that are both accurate and suitable for mobile deployment. By integrating advanced architectural features and efficient training techniques, this research contributes to the advancement of accessible and reliable skin cancer diagnostic tools, particularly in resource-limited and real-world environments.

3. MATERIALS & METHODS

3.1 Dataset Description

The **PAD-UFES-20** dataset is a thorough collection intended to facilitate the creation and assessment of deep learning models for skin cancer detection with non-dermoscopic pictures obtained through smartphone devices. This dataset includes **1,373 patients**, **1,641 unique skin lesions**, and **2,298 individual photos**. Although the sample size is smaller than that of existing datasets like HAM10000 or the ISIC archive, PAD-UFES-20 signifies a groundbreaking initiative in developing a publically accessible dataset designed for mobile-based skin lesion analysis.

A significant characteristic of PAD-UFES-20 is the variety of image acquisition conditions, as the photos are obtained utilizing different smartphone devices. This leads to considerable diversity in image resolution, dimensions, and lighting conditions, thereby accurately emulating real-world situations where image quality may be inconsistent. Each skin lesion in the collection is accompanied with a detailed set of **26 metadata attributes**, in addition to visual data. The attributes are classified into three primary categories: **patient clinical characteristics** (including age, gender, medical history, and lesion traits), **identifying features** (comprising patient ID, lesion ID, image ID, and biopsy confirmation status), and a **diagnostic label** denoting six types of skin lesions (ACK, BCC, SCC, MEL, NEV, and SEK).

The collection of clinical characteristics and skin lesion photos includes vital patient information, such as age and gender, along with relevant medical history factors including smoking status, alcohol intake, and prior cancer incidents. Furthermore, lesion-specific characteristics are thoroughly recorded, encompassing details such as lesion dimensions, anatomical position, and clinical presentations including pruritus, hemorrhage, and alterations in elevation.

3.2 Data Preprocessing

The PAD-UFES-20 dataset was subjected to an extensive preprocessing workflow to guarantee data integrity and enhance its suitability for deep learning model training.

Addressing Missing Values constituted the primary and essential phase in the preprocessing operation. Missing values for numerical variables, including Fitzpatrick skin type and lesion size, were imputed utilizing the K-Nearest Neighbors (KNN) imputer. This technique approximates absent values by calculating the mean of the nearest neighbors in the feature space, thus preserving the contextual significance of the data. In contrast, for categorical variables such as smoking status, alcohol use, and pesticide exposure, missing values were designated as “UNK” (Unknown). This method maintains the representation of uncertainty in the dataset without introducing bias, allowing the models to accurately identify and handle unknown or ambiguous information.

Subsequent to the imputation of absent values, **categorical variables** underwent transformation via one-hot encoding. This procedure transformed each categorical attribute into several binary columns, enabling their incorporation into machine learning models. The smoke attribute, categorized as False, True, and UNK, was divided into three separate columns: smoke_False, smoke_True, and smoke_UNK.

Normalization was conducted to continuous variables, including age and lesion size, using Z-score normalization to guarantee uniformity among numerical features. This method standardizes features to achieve a mean of zero and a standard deviation of one, mitigating numerical dominance from features with broader ranges and improving the convergence efficiency of deep learning models.

The dataset was further **divided into** training, validation, and test subsets via an **80-10-10 stratified** splitting method. Stratification guaranteed that the allocation of the target variable diagnostic categories remained uniform across all subgroups, thus preserving a balanced representation of each diagnostic class. The training set consisted of roughly 1,838 photos, the validation set included around 230 images, and the test set contained approximately 230 images. This stratified division was essential for alleviating class imbalance concerns and guaranteeing that the models were trained and assessed on representative samples of each class.

The final data preparation yielded a structured dataset comprising solely numerical features, amounting to **89 clinical features** following one-hot encoding and normalization. This meticulous preprocessing ensured that the dataset was well-suited for deep learning applications, enabling the effective integration of both visual and clinical data.

The preprocessing procedures implemented for the PAD-UFES-20 dataset created a solid basis for constructing precise and generalizable deep learning models for skin cancer detection utilizing non-dermoscopic smartphone photos. The

dataset was rigorously prepared for model training and evaluation by addressing missing values, encoding categorical variables, normalizing numerical features, and guaranteeing balanced data splits.

3.3 Overall Framework

This work introduces a **teacher–student** pipeline tailored for non-dermoscopic, smartphone-based skin lesion classification. The **teacher model** employs a ConvNeXt backbone, augmented with **Convolutional Block Attention Modules (CBAM)** to emphasize salient image regions, and integrates a **metadata encoder** that converts patient-specific attributes (e.g., age, risk factors) into high-dimensional embeddings. This metadata is then fused with the teacher's visually extracted features before passing through a **focal loss**–driven classification head, mitigating class imbalance.

Meanwhile, the **student model** adopts a more compact **EfficientNet-B0** backbone, also enhanced by CBAM and coupled with the same metadata encoding architecture. To bridge the performance gap between student and teacher, a **knowledge distillation** framework is employed, wherein the student acquires refined representations through soft logit alignment, feature matching, and attention-map transfer. This multi-term distillation objective allows the student to learn the teacher's rich, attention-driven feature space while maintaining a reduced parameter count—critical for deployment in low-resource or on-device environments. Through this synergy of **metadata integration**, **attention mechanisms**, and **knowledge distillation**, the proposed pipeline achieves robust skin lesion classification despite the variability and limited resolution of smartphone-acquired images.

3.4 Teacher Model

The teacher model (shown in utilizes a ConvNeXt architecture that incorporates a Convolutional Block Attention Module (CBAM) along with an additional metadata encoder. The ConvNeXt backbone enhances conventional convolutional neural networks by integrating innovative design features, such as high kernel sizes, depthwise separable convolutions, and refined layer normalization. This design seeks to leverage both sophisticated convolutional representations and supplementary clinical data.

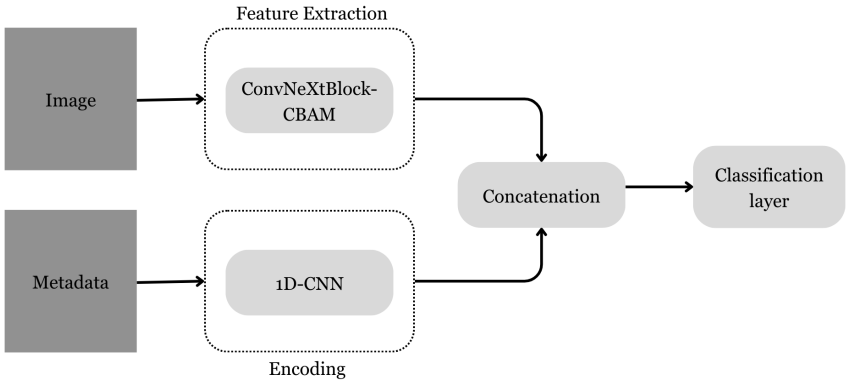


Figure 1. The overall framework of the teacher model

ConvNeXt Architecture. The ConvNeXt framework segments the network into several phases, each comprising a sequence of ConvNeXt blocks. The quantity of blocks and channel dimensions can be tailored; for example, the setup may utilize four stages with channel widths of 96, 192, 384, and 768, respectively. At each level, the features are downsampled and processed via a designated number of blocks (for instance, 3, 3, 9, 3), facilitating a progressive enhancement in representational depth. Each ConvNeXt block utilizes a depthwise convolution with a kernel size of 7 and a padding of 3 to efficiently capture local spatial patterns. The output of the depthwise convolution is subjected to layer normalization along the channel dimension to enhance training stability. Subsequently, two pointwise linear transformations are executed, the first of which increases the feature dimension (for example, by a factor of four), followed by a non-linear activation (e.g., GELU), and the second that reduces it back to its original size. A learnt scaling parameter γ, initialized at a minimal value (e.g., 1×10^{6}), regulates the block's residual output. This design enhances gradient flow and overall stability.

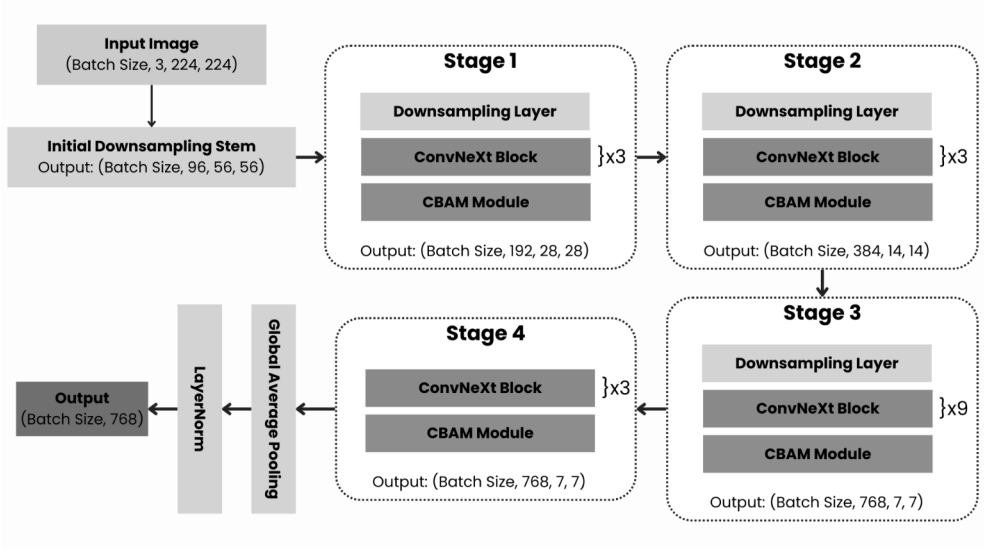


Figure 2. Image Encoder Architecture

Integration of CBAM. Following each stage, the output tensor is subjected to channel attention. This component aggregates features by global average pooling and global max pooling across spatial dimensions, resulting in two channel descriptors. Each descriptor undergoes processing by a common multi-layer perceptron prior to its amalgamation and subsequent passage through a sigmoid gating mechanism. Channels exhibiting greater relevance are thereby enhanced, whereas those with lesser discrimination are diminished. A spatial attention module is then utilized to enhance location-specific salience. The module calculates the average and maximum along the channel axis, concatenates these two feature maps, and subsequently processes them through a single convolutional layer followed by a sigmoid activation. Spatial attention accentuates prominent areas, enabling the model to concentrate on the lesion site in non-dermoscopic images, which frequently display background noise or inconsistent lighting conditions.

Metadata Encoder. The teacher model concurrently analyzes clinical or demographic data, known as metadata, alongside the primary visual pathway. The metadata generally comprises risk factors (e.g., familial cancer history, tobacco use), lesion characteristics (e.g., dimensions, site), and demographic information. An encoder, either a compact one-dimensional convolutional network or a multi-layer perceptron, translates the metadata into a feature vector with a dimensionality that corresponds to the final stage of ConvNeXt (e.g., 768). Batch normalization and non-linear activations, such as ReLU, are utilized to maintain stability and enhance expressive capacity.

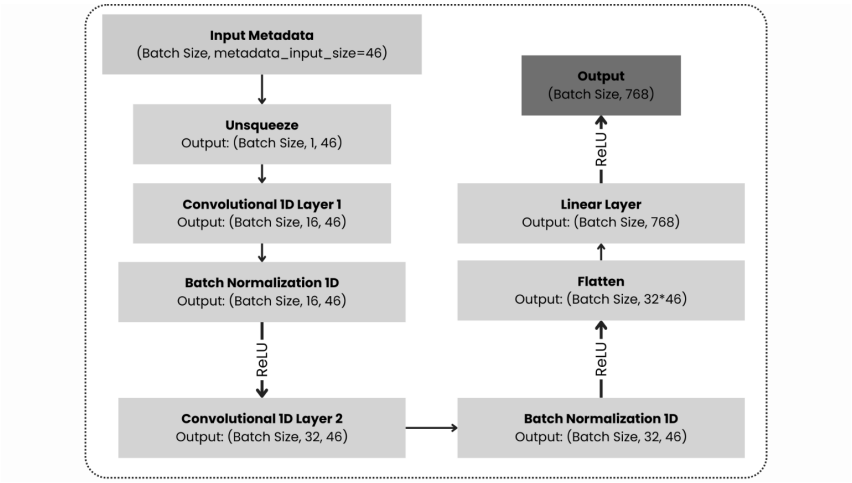


Figure 3. Metadata Encoder

Data Integration and Categorization. Following the processing of input images via the ConvNeXt backbone augmented with Convolutional Block Attention Modules (CBAM), a global average pooling operation is executed to diminish the spatial dimensions, yielding a visual feature vector $v \in R^{768}$. The corresponding metadata is encoded into a feature vector $m \in R^{768}$ using a specialized metadata encoder. The two vectors are concatenated to create a fused feature vector f, mathematically represented as: $f = v \parallel m \in R^{1536}$

The fused vector f is subsequently passed through a multi-layer perceptron (MLP) comprising a fully connected layer with ReLU activation and dropout regularization, followed by a final linear layer that outputs the class logits $z \in R^C$, where CCC is the number of diagnostic classes. This classification head can be represented as:

$$z = \text{Linear}_2\left(\text{Dropout}\left(\text{ReLU}\left(\text{Linear}_1(f)\right)\right)\right)$$

The teacher model integrates visual features with metadata, utilizing both image-based and contextual clinical information to improve its accuracy in classifying skin lesions. This fusion method guarantees that the model gains a thorough representation of each sample, resulting in more discerning and dependable predictions across various real-world situations.

3.5 Student Model

The student model is designed to operate under constrained computational environments while preserving high classification accuracy through knowledge distillation. Its foundation is an **EfficientNet-B0** backbone, selected for its demonstrated balance between accuracy and efficiency. By applying compound scaling principles, EfficientNet-B0 systematically adjusts network depth, width, and resolution, thereby reducing the parameter count and memory footprint relative to larger architectures key advantages for on-device deployment in settings with limited resources.

To bolster the student network's representational capacity, **Convolutional Block Attention Modules (CBAM)** are integrated into its feature extraction pipeline. In practice, the code leverages the timm library to instantiate efficientnet_b0 with features_only=True, which returns the intermediate feature maps at four predefined indices. Each of these feature maps is then passed through a dedicated CBAM module, where **channel attention** re-weights the feature channels according to their relevance, and **spatial attention** refines important spatial regions within each map. This ensures that despite being smaller, the student model can still learn to focus on clinically salient regions—an ability that the teacher also exploits via CBAM.

Alongside the image-based feature extractor, the student mirrors the teacher's **metadata encoder** to incorporate non-visual patient data. The encoder operates on numerical and one-hot–encoded variables (e.g., age, smoking habits, family cancer history) using a series of 1D convolutional layers, interspersed with batch normalization and ReLU activations, ultimately producing a fixed-length embedding (e.g., 768 dimensions). This embedding is designed to capture important clinical indicators that might not be visibly apparent in the image alone.

Once the image features and metadata embedding have been computed, they are concatenated to form a fused vector. To facilitate **feature-level knowledge distillation**, a **projection layer** (implemented as a single linear transformation) then maps this fused representation to the teacher's embedding dimension (e.g., 1536). This step is crucial for aligning the student's final features with those of the teacher, thereby allowing a direct application of mean squared error (MSE) during the distillation process. If the dimensions are mismatched, meaningful feature comparisons become significantly more challenging.

Finally, the projected representation enters a **classification head**, which comprises two linear layers (with dropout and ReLU activation) that generate logits across the six skin lesion classes. Despite employing fewer parameters than the teacher, the student thereby preserves the ability to differentiate among various lesion types. Through knowledge distillation incorporating logit alignment, feature matching, and attention transfer the student acquires representations and attention patterns akin to the teacher's, all while maintaining the compactness required for low-resource deployment.

3.6 Knowledge Distillation Setup and Implementation

The objective of knowledge distillation is to train the student network to approximate the teacher's performance by learning from the teacher's richer outputs. In this work, distillation incorporates multiple loss components classification, logit alignment, feature matching, and attention transfer to capture both **final** and **intermediate** representations of the

teacher model. This integrated approach equips the student with comparable diagnostic power while preserving a lighter footprint.

Classification Term (Focal Loss)

Given the class imbalance often observed in skin lesion datasets, a **Focal Loss** replaces the standard cross-entropy loss. Focal Loss applies a focusing parameter γ that amplifies the gradient signal for difficult or minority samples, thus mitigating the dominance of common classes. Additionally, class weights (computed through compute_class_weight and fine-tuned for certain rare or critical classes) are incorporated into the loss function. This ensures that the student prioritizes underrepresented lesion types during training, improving recall for malignancies.

Logit-Based Distillation (KL Divergence)

Traditional knowledge distillation begins with comparing the **softened logits** of the teacher and student via a **Kullback Leibler (KL) divergence** term. To obtain these softened logits, both the teacher and student outputs are divided by a **temperature** T. As T rises above 1, the output probability distributions become smoother, highlighting the relative relationships between classes. Minimizing KL divergence encourages the student to mimic the teacher's class probability structure, guiding it to discriminate among lesion classes in a teacher-like manner.

Feature Matching (MSE)

In addition to matching the final prediction layer, the student is prompted to align its **deep feature representation** with that of the teacher. Specifically, the mean squared error (MSE) is computed between the teacher's and student's **fused embeddings** (i.e., the concatenated image and metadata features). To enable a direct comparison, the student's projection layer ensures the dimensionality of the student features aligns with the teacher's feature space (e.g., 1536 dimensions). By enforcing a feature-level similarity, the student not only learns the final classification distribution but also converges toward the teacher's high-level representation of image and metadata cues.

Attention Transfer (MSE)

Both teacher and student networks employ **Convolutional Block Attention Modules (CBAM)** that produce **spatial attention** maps for each intermediate feature stage. These attention maps indicate where (and how strongly) the network attends to specific regions in the input. To encourage the student to focus on the same lesion regions the teacher deems important, an MSE loss is applied between corresponding teacher and student attention maps. If there is a dimension mismatch (e.g., differing feature map sizes), the teacher's maps are bilinearly interpolated to match the student's resolution prior to comparison. This step promotes consistent spatial reasoning in both networks, benefiting the student's interpretability and potentially improving detection of subtle lesion features.

Weighted Loss Combination

The total distillation loss is composed of these four terms:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{Focal}} + \alpha\,\mathcal{L}_{\text{KD}} + \beta\,\mathcal{L}_{\text{Feature}} + \gamma\,\mathcal{L}_{\text{Attention}}$$

where:

- L_{Focal} is the focal classification loss.
- L_{KD} is the KL divergence (teacher–student logit matching).
- $L_{Feature}$ is the feature MSE.
- $L_{Attention}$ is the attention map MSE.
- α, β, and γ are tunable hyperparameters that balance the contributions of the different distillation components.

By adjusting these coefficients, the training can be tailored to emphasize classification accuracy, better mimic the teacher's intermediate representations, or align attention patterns more closely. In practice, careful tuning on the validation set ensures that the student leverages the teacher's guidance without overly constraining its own learning capacity. Through this framework, the student achieves strong performance with fewer parameters, offering a compelling solution for **smartphone-based** or **resource-limited** skin lesion classification scenarios.

4. EXPERIMENTAL SETUP

Training Configuration. All experiments are implemented in PyTorch to leverage GPU acceleration where available. For efficient parallelization, the student model may be wrapped in a DataParallel module if multiple GPUs are present, while the teacher operates in inference mode to provide static outputs for knowledge distillation. A moderate batch size (for instance, 32) is selected to balance hardware memory constraints and robust gradient estimation.

Both the teacher and student models use the **AdamW** optimizer, which applies weight decay of 10^4 to counter overfitting. The learning rate is scheduled via a **CosineAnnealingLR**, gradually decreasing over epochs to encourage stable convergence. The teacher model is trained first, using focal loss to manage class imbalance. Once validated, its parameters are frozen and saved for subsequent distillation. During student training, each mini-batch is processed by both teacher and student, enabling the student to integrate the teacher's knowledge—encompassing logits, features, and attention maps—into its own parameter updates.

Early Stopping Strategy. The student's progress is monitored using both validation loss and performance metrics such as weighted F1-score. A patience threshold is set so that if no improvement is observed within a fixed number of epochs, a cooldown phase begins. If the model continues to show no progress after the cooldown, training is terminated. This mechanism prevents overfitting by halting training at a point where additional epochs would not significantly improve generalization.

Evaluation Protocol. Following training, the best-performing student checkpoint—identified by the lowest validation loss or highest F1-score—is evaluated on a held-out test set drawn from PAD-UFES-20. The primary metrics include overall accuracy, precision, recall, and F1-score for each class. A confusion matrix offers a detailed view of misclassifications, highlighting whether the student struggles with specific lesion types. Since the dataset involves a non-dermoscopic setting, these per-class analyses can illuminate how well the student copes with image noise and variability in lesion appearance.

5. RESULTS & DISCUSSION

5.1 Teacher Model Performance

The training phase for the proposed teacher model demonstrated a consistent trend toward convergence. Both the training and validation accuracy increased steadily during the initial epochs. Eventually, the training accuracy surpassed 99%, while the validation accuracy improved in parallel until approximately the 60th epoch. Beyond that point, the widening gap between training and validation accuracy indicated that the model began overfitting to the training data. Nonetheless, the validation accuracy plateaued at about 80%, suggesting the model was not underfitting but had instead reached a performance limit. When evaluated on the unseen test set, the teacher model achieved an overall accuracy of **80.43%**, a noteworthy result for non-dermoscopic, smartphone-based skin lesion classification. Given the class imbalance in the PAD-UFES-20 dataset, it is also important to consider the **F1-score**, which combines both precision and recall. The teacher model's F1-score of **80.46%** demonstrates a balanced performance in handling false positives and false negatives, reflecting its capacity to discern challenging lesion categories effectively.

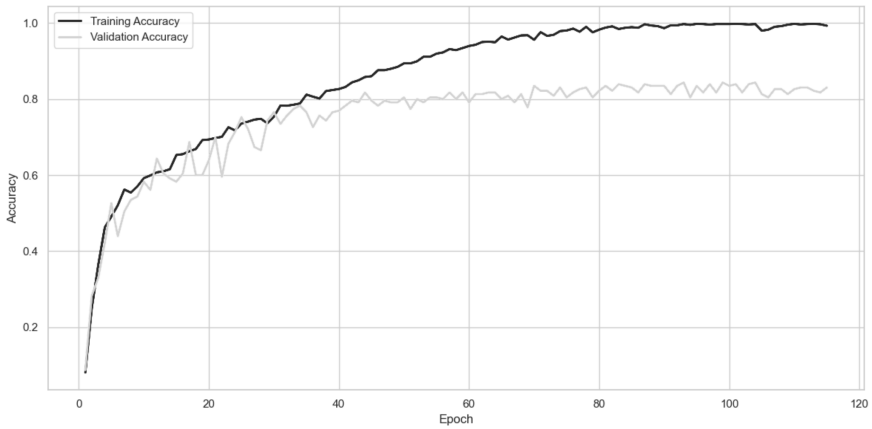


Figure 4. Comparison of the training and validation accuracy across epochs.

A closer inspection of the teacher's confusion matrix reveals that while the model correctly identifies the majority of class ACK and class BCC instances, it occasionally confuses these with each other, indicating partial overlap in lesion appearance or metadata patterns. Certain misclassifications within minority classes (e.g., class SCC) further highlight the difficulty in distinguishing lesions that lack sufficient training examples or present visually ambiguous features. Nevertheless, the teacher model's use of **Convolutional Block Attention Modules (CBAM)** and metadata appears to systematically reduce gross errors across most categories.

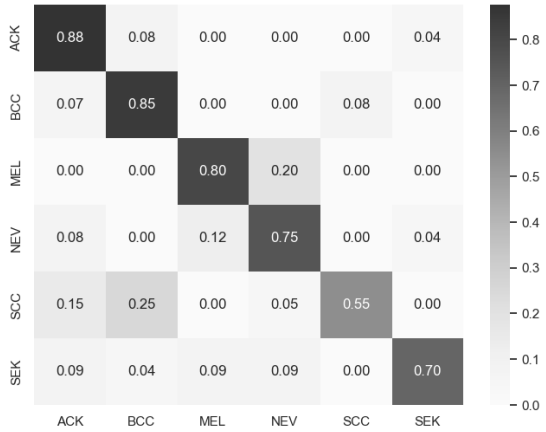


Figure 5. Confusion Matrix (Probability).

Two notable frameworks that incorporate metadata alongside image-based features are **MetaBlock** and **DeepMetaForge**. **MetaBlock** (proposed by A. G. C. Pacheco and R. A. Krohling [16]) refines the most salient features extracted from lesion images through an adaptive metadata-driven mechanism, outperforming other feature-fusion approaches in multiple scenarios. **DeepMetaForge** (developed by J. Wang et al. [24]) employs a vision transformer backbone (BEiT) and then integrates the encoded metadata with visual representations in a compressed format. Their empirical tests, conducted on various dermoscopic and smartphone-acquired datasets, underscore the viability of unifying image data with demographic or clinical attributes for more robust diagnosis. In our study, we compare both methods—MetaBlock and DeepMetaForge—to our proposed model to contextualize performance gains and identify the advantages of different metadata-integration strategies.

In comparison to other approaches—specifically MetaBlock and DeepMetaForge—the teacher model surpassed both in overall accuracy (80.43%) and F1-score (80.46%), as summarized in **Table 1**. All methods were evaluated on the same preprocessed dataset under similar conditions, integrating both image features and patient metadata. The improved

results suggest that the combination of a **convolutional attention mechanism** (CBAM) with non-dermoscopic images, supplemented by relevant demographic or clinical information, can enhance diagnostic reliability.

Table 1. Comparison of Proposed Model to the existing approaches

Comparison of Proposed Model				
Model	**Accuracy**	**Precision**	**Recall**	**F1-score**
ResNet50-MetaBlock	0.7261	0.6766	0.7261	0.6969
BEiT-DeepMetaForge	0.7696	0.7726	0.7696	0.7695
Proposed Model	**0.8043**	**0.8088**	**0.8043**	**0.8046**

Beyond simply outperforming prior approaches, the teacher model highlights the importance of **patient metadata** in classification. As shown in **Table 2**, even when using an advanced backbone architecture (such as ConvNeXt), performance can remain limited if metadata features are excluded or if the attention mechanism is removed. Under constraints like resized image inputs or normalized scales, models may struggle to extract critical lesion details solely from raw images. Integrating metadata into the attention-based convolutional architecture proved to significantly bolster diagnostic robustness, demonstrating that multi-modal data fusion can surpass single-modality image-based strategies in tackling skin lesions.

Table 2. The Influence of Metadata Encoder and Attention Mechanism On The Proposed Model.

The Influence of Metadata Encoder and Attention Mechanism				
Model	**Accuracy**	**Precision**	**Recall**	**F1-score**
Proposed Model	**0.8043**	**0.8088**	**0.8043**	**0.8046**
Without Metadata Encoder	0.6174	0.5784	0.6174	0.5914
Without Both Metadata Encoder & Attention Mechanism	0.4435	0.4675	0.4435	0.4514

Lastly, **Table 3** underscores the impact of using **Focal Loss** for handling class imbalance, showing marked improvements in F1-scores across nearly all lesion categories. For example, ACK rose from 0.8077 to 0.8533, and NEV increased from 0.7234 to 0.7826, indicating that Focal Loss helps the model allocate greater learning capacity to classes that are more difficult or occur less frequently.

Table 3. Improvement In F1-Scores by Using Focal Loss

Improvement In F1-Scores by Using Focal Loss		
Class	**Without Focal Loss**	**With Focal Loss**
ACK	0.8077	0.8533
BCC	0.8466	0.8521
MEL	0.5000	0.5714
NEV	0.7234	0.7826
SCC	0.5854	0.5989
SEK	0.6341	0.7442

5.2 Student Model Performance

Comparison with the Teacher. Despite having significantly fewer parameters, the student model ultimately matches the teacher's **80.43%** test accuracy, achieving a weighted F1-score of **0.8016**. This result underscores the efficacy of the **knowledge distillation** framework, wherein the student inherits not only the teacher's soft logits via KL divergence but also intermediate feature representations and attention patterns. Hence, the smaller EfficientNet-B0 network preserves the teacher's diagnostic strengths while offering enhanced computational efficiency.

Table 4. Performance of Both Teacher and Student Models

Performance Table Comparision				
Model	**Accuracy**	**Precision**	**Recall**	**F1-score**
Teacher Model	0.8043	0.8088	0.8043	0.8046
Student Model	0.8043	0.8053	0.8043	0.8016

Per-Class Observations. The student model's performance shows a pattern broadly similar to the teacher's, with strong performance in classes ACK and BCC and moderate challenges in class SCC. Interestingly, the student displays an encouraging recall of 0.80 for class MEL, despite class MEL's limited sample size. This outcome points to the tangible benefits of transferring the teacher's attention maps, which highlight pertinent lesion regions that might otherwise be overlooked by a compact network.

Table 5. F1-score Illustration of Student and Teacher Models

F1-score Illustration		
Class	**Teacher Model**	**Student Model**
ACK	0.8533	0.8228
BCC	0.8521	0.8434
MEL	0.5714	0.7273
NEV	0.7826	0.8511
SCC	0.5789	0.5714
SEK	0.7442	0.7442

Training Behavior. While the student starts with a higher training loss (values above 5.0), it converges steadily with each epoch, guided by the teacher's refined outputs. The classification metrics confirm that the student's validation accuracy and F1-scores eventually align with those of the teacher, even though the absolute training loss does not reach as low a value as the teacher's. This observation suggests that the student does not need to replicate every nuance of the teacher's parameters to achieve comparable predictive performance.

5.3 Discussion

This study underscores the potential of a multi-faceted knowledge distillation strategy for improving skin lesion classification when only non-dermoscopic, smartphone-acquired images are available. By transferring not just the teacher model's logits but also its intermediate features and attention maps, the proposed approach enables a lightweight EfficientNet-B0–based student to approximate the performance of a more complex, ConvNeXt-based teacher. Crucially, both models integrate metadata (e.g., demographic and risk factors) to enhance discrimination, particularly for underrepresented lesion types. Below, we elaborate on the principal contributions of this work, discuss key limitations, and propose avenues for future investigation.

From a methodological standpoint, **the first major contribution** is the demonstration that logit-based distillation, combined with feature matching and attention-transfer losses, imparts a compact network with teacher-level representational quality. This configuration ensures that the student inherits not only the final prediction distribution but also deeper, spatially grounded cues from the teacher's CBAM-enhanced feature maps. **Second,** by incorporating diverse

metadata alongside image features in both teacher and student, the study highlights how patient demographics and behavioral factors can refine the model's decision boundaries, mitigating noise and variability inherent to smartphone-captured images. **Third,** the near parity of the student's accuracy and weighted F1-score with those of the teacher validates the resource-efficiency of the distilled model—a key asset for on-device deployment in resource-constrained or remote healthcare settings.

Despite these promising outcomes, the proposed framework does exhibit certain **limitations**. **First,** class imbalance remains a prominent obstacle. Although focal loss and targeted class-weight adjustments helped address minority lesion categories, certain classes (such as class 4) retained comparatively low recall. This shortfall underlines the ongoing necessity for more specialized strategies, whether through advanced oversampling, refined augmentation, or domain-specific data enrichment. **Second,** the dataset used—PAD-UFES-20—while valuable for representing non-dermoscopic conditions, may not capture the full breadth of dermatological diversity across varying geographic or ethnic populations. Such potential dataset bias could constrain the model's generalizability and warrant additional validations on different cohorts. **Third,** although the teacher model's training did not exhibit extreme overfitting, its parameter size and extremely low final training loss imply that further regularization may be necessary under more stringent or smaller datasets. **Fourth,** while CBAM-based attention maps introduce a measure of interpretability, truly integrating these insights into a clinical pipeline would likely demand broader stakeholder engagement, including dermatological experts, to translate visual attention cues into tangible diagnostic confidence.

Based on these observations, there are several **future directions**. **To begin with,** further balancing measures—including sophisticated data augmentation approaches that specifically target rare lesion morphologies—could complement focal loss in raising minority-class performance. **In addition,** exploring a richer set of clinical variables and advanced metadata-fusion mechanisms may unlock subtle diagnostic patterns beyond those discernible from images alone. **Cross-dataset evaluation** likewise represents a crucial milestone for validating transferability: repeating these experiments on other non-dermoscopic data sources would help to confirm the robustness of both teacher and student architectures. Finally, **further compression** of the student network, whether through quantization or pruning, could facilitate real-time inference on even modest smartphone processors, thereby broadening the scope for low-cost, widespread adoption in remote clinical settings.

Taken together, these findings bolster confidence in the use of metadata-augmented, attention-driven knowledge distillation for smartphone-based lesion classification. The synergy of carefully aligned logit, feature, and attention-transfer losses emerges as a potent way to preserve diagnostic fidelity in a compact model. While additional refinements are undoubtedly required to address class imbalance and dataset limitations, the overarching framework makes a convincing case for the feasibility of low-resource, high-accuracy dermatological screening.

6. CONCLUSION

This study has demonstrated the feasibility of a **knowledge distillation**–based pipeline for **skin lesion classification** using **non-dermoscopic smartphone images**, targeting scenarios where computational resources and specialized diagnostic tools are limited. By transferring learned representations from a larger, metadata-aware ConvNeXt teacher model to an EfficientNet-B0 student, the proposed framework manages to preserve high classification accuracy and robust attention mapping while substantially reducing inference complexity. Incorporating **Convolutional Block Attention Modules (CBAM)** and leveraging **focal loss** further bolster the system's ability to handle imbalanced lesion classes, thus allowing more clinically significant yet underrepresented categories to be recognized effectively.

More specifically, the **teacher–student** arrangement capitalizes on **logit matching**, **feature alignment**, and **attention transfer** to guide the student model toward the teacher's nuanced feature representations. The additional **metadata integration** refines decision boundaries by contextualizing lesion appearance with patient-specific factors, underlining the benefit of fusing tabular clinical information with image data. As demonstrated with the PAD-UFES-20 dataset, this synergy ultimately supports near teacher-level performance on six distinct lesion types.

In settings where dermoscopic imaging devices or expert personnel are unavailable, this solution is well positioned to deliver accurate lesion screening directly on smartphones. By reducing model size and computational requirements without compromising diagnostic quality, it opens possibilities for earlier intervention and improved access to

dermatological assessments in rural or resource-constrained communities. Consequently, the proposed method not only strengthens the case for **knowledge distillation** in medical imaging but also encourages broader adoption of **metadata-fused**, **attention-driven** architectures for real-world, **mobile health** applications.

REFERENCES

[1] Z. Apalla, D. Nashan, R. B. Weller, et al., "Skin cancer: epidemiology, disease burden, pathophysiology, diagnosis, and therapeutic approaches," *Dermatol Ther (Heidelb)*, vol. 7 (Suppl 1), pp. 5–19, 2017.

[2] M. Zafar, M. I. Sharif, M. I. Sharif, S. Kadry, S. A. C. Bukhari, and H. T. Rauf, "Skin lesion analysis and cancer detection based on machine/deep learning techniques: a comprehensive survey," *Life*, vol. 13, p. 146, 2023.

[3] V. Subbhuraam and D. Panigrahi, "Telemedicine," in *Predictive Analytics in Healthcare, Volume 1*, IOP Publishing, 2021, pp. 4-1–4-37, doi: 10.1088/978-0-7503-2312-3ch4.

[4] Zhang, B., Zhou, X., Luo, Y. et al. Opportunities and Challenges: Classification of Skin Disease Based on Deep Learning. Chin. J. Mech. Eng. 34, 112 (2021). https://doi.org/10.1186/s10033-021-00629-5

[5] M. A. Kassem, K. M. Hosny, R. Damaševičius, and M. M. Eltoukhy, "Machine Learning and Deep Learning Methods for Skin Lesion Classification and Diagnosis: A Systematic Review," Diagnostics, vol. 11, no. 8, p. 1390, Jul. 2021, doi: 10.3390/diagnostics11081390.

[6] M. A. Sabri, Y. Filali, S. Fathi, and A. Aarab, "Detection, Analysis and Classification of Skin Lesions: Challenges and Opportunities," in Intelligent Systems in Big Data, Semantic Web and Machine Learning, N. Gherabi and J. Kacprzyk, Eds. Cham: Springer International Publishing, 2021, pp. 197–214. doi: 10.1007/978-3-030-72588-4_14.

[7] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," arXiv preprint arXiv:1503.02531, 2015. [Online]. Available: https://arxiv.org/abs/1503.02531.

[8] A. G. Pacheco, G. R. Lima, A. S. Salomão, B. Krohling, I. P. Biral, G. G. De Angelo, F. C. Alves Jr, J. G. Esgario, A. C. Simora, P. B. Castro, F. B. Rodrigues, P. H. Frasson, R. A. Krohling, H. Knidel, M. C. Santos, R. B. Do Espírito Santo, T. L. Macedo, T. R. Canuto, and L. F. De Barros, "PAD-UFES-20: A skin lesion dataset composed of patient data and clinical images collected from smartphones," *Data in Brief*, vol. 32, p. 106221, 2020, doi: 10.1016/j.dib.2020.106221.

[9] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A ConvNet for the 2020s," arXiv preprint arXiv:2201.03545, 2022. [Online]. Available: https://arxiv.org/abs/2201.03545.

[10] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," arXiv preprint arXiv:1807.06521, 2018. [Online]. Available: https://arxiv.org/abs/1807.06521.

[11] M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," arXiv preprint arXiv:1905.11946, 2020. [Online]. Available: https://arxiv.org/abs/1905.11946.

[12] P. Yao *et al*., "Single model deep learning on imbalanced small datasets for skin lesion classification," *IEEE Transactions on Medical Imaging*, vol. 41, no. 5, pp. 1242–1254, May 2022, doi: 10.1109/TMI.2021.3136682.

[13] L. Hoang, S.-H. Lee, E.-J. Lee, and K.-R. Kwon, "Multiclass skin lesion classification using a novel lightweight deep learning framework for smart healthcare," *Applied Sciences*, vol. 12, p. 2677, 2022, doi: 10.3390/app12052677.

[14] S. Benyahia, B. Meftah, and O. Lézoray, "Multi-features extraction based on deep learning for skin lesion classification," *Tissue and Cell*, vol. 74, p. 101701, 2022, doi: 10.1016/j.tice.2021.101701.

[15] S. K. Datta, M. A. Shaikh, S. N. Srihari, and M. Gao, "Soft attention improves skin cancer classification performance," in *Interpretability of Machine Intelligence in Medical Image Computing, and Topological Data Analysis and Its Applications for Medical Data. IMIMIC TDA4MedicalData 2021 2021*, M. Reyes *et al.* (Eds.), Lecture Notes in Computer Science, vol. 12929, Springer, Cham, 2021, doi: 10.1007/978-3-030-87444-5_2.

[16] A. G. C. Pacheco and R. A. Krohling, "An attention-based mechanism to combine images and metadata in deep learning models applied to skin cancer classification," *IEEE Journal of Biomedical and Health Informatics*, vol. 25, no. 9, pp. 3554–3563, Sept. 2021, doi: 10.1109/JBHI.2021.3062002.

[17] C. Dong, D. Dai, Y. Zhang, C. Zhang, Z. Li, and S. Xu, "Learning from dermoscopic images in association with clinical metadata for skin lesion segmentation and classification," Computers in Biology and Medicine, vol. 152, p. 106321, 2022, doi: 10.1016/j.compbiomed.2022.106321.

[18] A. Pundhir, S. Dadhich, A. Agarwal, and B. Raman, "Towards improved skin lesion classification using metadata supervision," in 2022 26th International Conference on Pattern Recognition (ICPR), Montreal, QC, Canada, 2022, pp. 4313–4320, doi: 10.1109/ICPR56361.2022.9956071.

[19] A. Dascalu, B. N. Walker, Y. Oron, *et al*., "Non-melanoma skin cancer diagnosis: a comparison between dermoscopic and smartphone images by unified visual and sonification deep learning algorithms," *Journal of Cancer Research and Clinical Oncology*, vol. 148, pp. 2497–2505, 2022, doi: 10.1007/s00432-021-03809-x.

[20] Y. Wang, Y. Wang, J. Cai, T. K. Lee, C. Miao, and Z. J. Wang, "SSD-KD: A self-supervised diverse knowledge distillation method for lightweight skin lesion classification using dermoscopic images," *Medical Image Analysis*, vol. 84, p. 102693, 2023, doi: 10.1016/j.media.2022.102693.

[21] 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, 2023, doi: 10.1016/j.compbiomed.2023.106571.

[22] M. R. Kabir, R. H. Borshon, M. K. Wasi, R. M. Sultan, A. Hossain, and R. Khan, "Skin cancer detection using lightweight model souping and ensembling knowledge distillation for memory-constrained devices," *Intelligence-Based Medicine*, vol. 10, p. 100176, 2023, doi: 10.1016/j.ibmed.2024.100176.

[23] R. Yadav and A. Bhat, "Skin cancer detection using knowledge distillation," in *2024 3rd International Conference for Innovation in Technology (INOCON)*, Bangalore, India, 2024, pp. 1–4, doi: 10.1109/INOCON60754.2024.10512006.

[24] S. Vachmanus, T. Noraset, W. Piyanonpong, T. Rattananukrom, and S. Tuarob, "DeepMetaForge: A deep vision-transformer metadata-fusion network for automatic skin lesion classification," *IEEE Access*, vol. 11, pp. 145467–145484, 2023, doi: 10.1109/ACCESS.2023.3345225.