

Skin Cancer Detection Using Deep Learning

M. Pravallika¹, MD. Kouser Ali², M. Dinesh Kumar³, P. Sai Mehar⁴

¹ Student, Dept. of Electronics and Communication Engineering, Seshadri Rao Gudlavalleru Engineering College, Gudlavalleru, Andhra Pradesh, India

² Student, Dept. of Electronics and Communication Engineering, Seshadri Rao Gudlavalleru Engineering College, Gudlavalleru, Andhra Pradesh, India

³ Student, Dept. of Electronics and Communication Engineering, Seshadri Rao Gudlavalleru Engineering College, Gudlavalleru, Andhra Pradesh, India

⁴ Student, Dept. of Electronics and Communication Engineering, Seshadri Rao Gudlavalleru Engineering College, Gudlavalleru, Andhra Pradesh, India

Abstract - Skin disease, especially melanoma, is among the most dangerous health conditions worldwide. Early detection is useful for successful treatment. However, diagnosis of skin lesions requires dermatological expertise, which may be unavailable in rural or resource-limited regions, and can be highly time-consuming. There is a gap that can be fulfilled by accurate and accessible diagnostic support systems. A computer-aided detection tool can assist clinicians by providing a fast, reliable second opinion, reducing delays, and increasing diagnostic consistency for skin cancer detection. This project implements a deep learning-based skin cancer classification system that is used for the classification of melanoma and benign classes using convolutional neural network (CNN) models, namely MobileNetV2_S and EfficientNetB3. It is implemented in TensorFlow through transfer learning. Skin lesion images taken from the ISIC 2020 dataset are resized and normalized before being fed to the CNN models for classification into benign and melanoma. CNNs' models are used for feature extraction, which learns spatial lesion patterns. The proposed method also includes data augmentation, class balancing, fine-tuning, threshold optimization, hyperparameters, and test time augmentation to improve classification performance. The performance of the models is evaluated by using metrics such as accuracy, precision, recall, specificity, F1-score, and AUC. These experimental results show EfficientNetB3 achieved the best overall test performance with 94.72% accuracy, 96.26% specificity, 94.06% melanoma F1-score, and 0.9860 AUC, while MobileNetV2_S provided competitive performance with 92.24% accuracy and lower computational cost, which results in low training time.

Key Words: melanoma, deep learning, transfer learning, computer-aided diagnosis, MobileNetV2, EfficientNetB3

1. INTRODUCTION

Skin cancer is one of the most common forms of cancer, and melanoma is clinically significant because of its high metastatic potential when diagnosis is delayed. Visual examination and dermoscopy remain central to routine

screening, but interpretation is influenced by clinical experience, lesion diversity, imaging variability, and time constraints. As a result, automated systems that can analyze dermoscopic images consistently and rapidly are increasingly valuable as decision-support tools [1], [2].

Recent advances in deep learning have transformed medical image analysis by allowing convolutional neural networks to learn discriminative patterns directly from images. In dermatology, these networks can capture texture, pigment distribution, asymmetry, border irregularity, and other lesion characteristics that are difficult to encode manually. Public datasets such as ISIC have accelerated this research by providing large-scale dermoscopic collections for melanoma analysis [3], [7].

The objective of this work is to develop and compare two transfer-learning based binary classifiers, MobileNetV2_S and EfficientNetB3, for skin lesion classification into benign and melanoma categories. The implementation emphasizes practical deployment features: lightweight preprocessing, data augmentation, class balancing, two-stage fine-tuning, threshold optimization constrained by specificity, and test-time augmentation. The main contributions of the paper are: (i) a reproducible TensorFlow-based binary classification pipeline for ISIC2020 images, (ii) a detailed comparison between a lightweight and a stronger backbone under the same training protocol, and (iii) quantitative analysis of classification quality and inference cost for clinician-assistance scenarios.

2. RELATED WORK

Existing literature shows clear progression from handcrafted feature pipelines to end-to-end deep neural networks for melanoma detection. The review by Kaur et al. highlights that high-performing computer-aided diagnosis systems often combine careful preprocessing, lesion-focused analysis, and deep classification models to improve melanoma discrimination [1]. Their study also underlines the relevance of ISIC2020 as a challenging benchmark for modern melanoma CAD research.

Practical systems have used pretrained CNN backbones such as VGG-based networks and MobileNet to enable accessible skin cancer screening. Mohankumar et al. presented a web-oriented diagnostic workflow for benign and malignant classification and showed that transfer learning can support real-time clinician assistance with satisfactory validation performance [2]. These studies confirm that transfer learning is effective even when medical datasets are smaller than large natural-image corpora.

Dataset quality and evaluation design remain equally important. Rotemberg et al. described the patient-centric SIIM-ISIC 2020 dataset and its value for melanoma research in clinically meaningful settings [3]. Cassidy et al. later showed that ISIC image collections contain duplicate and near-duplicate samples that can bias evaluation if not handled carefully [7]. Motivated by these observations, the present work focuses on a controlled split-based evaluation with threshold tuning and robust test-time averaging, while comparing MobileNetV2 [4] and EfficientNet [5] under identical experimental conditions.

3. MATERIALS AND METHODS

3.1 Dataset and Experimental Setup

The implementation uses the ISIC2020 dermoscopic image collection organized into train, validation, and test folders. After extraction, the training pipeline processed 9,017 images for training, 1,287 images for validation, and 2,578 images for final testing. The task is binary classification, where class label 0 corresponds to benign lesions, and class label 1 corresponds to melanoma. The test-set confusion matrices show 1,419 benign and 1,159 melanoma images, indicating that the evaluation set remains clinically meaningful for both classes.

All images were resized to $224 \times 224 \times 3$. Backbone-specific preprocessing from TensorFlow Keras applications was applied before model input. The training generator used moderate augmentation to improve generalization: rotation up to 20 degrees, width and height shifts of 0.03, zoom up to 0.08, horizontal flipping, and nearest-neighbor filling for empty pixels.

3.2 Transfer-Learning Models

Two pretrained CNN backbones were investigated. The first was MobileNetV2_S, implemented with MobileNetV2 and width multiplier $\alpha = 0.75$ to reduce computational cost. The second was EfficientNetB3, selected as a stronger backbone with higher representational capacity. Both models were initialized with ImageNet weights and used global average pooling at the backbone output.

A common custom classification head was attached to each backbone. This head consisted of batch normalization, a dropout layer of 0.35, a fully connected layer with 512 units, batch normalization, LeakyReLU activation, a dropout layer of 0.40, a second dense layer

with 256 units, batch normalization, LeakyReLU activation, a dropout layer of 0.30, and a final sigmoid classifier for binary prediction. The total parameter count was approximately 2.18 million for MobileNetV2_S and 11.71 million for EfficientNetB3.

3.3 Training Strategy

Training was performed in two phases. In Phase 1, the backbone was frozen and only the custom classifier head was trained for 15 epochs. In Phase 2, the top 30% of the backbone layers, excluding batch-normalization layers, were unfrozen for fine-tuning with an additional schedule of up to 35 epochs. The Adam optimizer was used with learning rate $1e-3$ in Phase 1 and $1e-5$ during fine-tuning.

Binary cross-entropy was used as the loss function. Accuracy, precision, recall, and AUC were monitored during training. To handle class imbalance, balanced class weights were computed from the training labels, resulting in weights of 0.9081 for benign and 1.1127 for melanoma. Model Checkpoint, Reduce LROnPlateau, and Early Stopping callbacks were configured to maximize validation of AUC and prevent overfitting.

3.4 Threshold Optimization and Test-Time Augmentation

Instead of using a fixed decision threshold of 0.50 for all predictions, the validation set probabilities were searched over thresholds from 0.10 to 0.90 in 161 steps. Only thresholds satisfying a minimum validation specificity of 0.95 were considered, and the final threshold was selected using the highest F1 score within that feasible set. This design favors lower false-positive rates during tuning; however, varying generalization on the unseen test set resulted in final test specificities of 91.05% for MobileNetV2_S and 96.26% for EfficientNetB3.

For final inference, a five-pass test-time augmentation was used. Each test image was evaluated once in its original form and four additional times with horizontal flipping, rotation up to 10 degrees, and zoom up to 0.05. The probabilities from all passes were averaged to obtain a more stable final score.

3.5 Evaluation Metrics

The models were evaluated using accuracy, weighted precision, weighted recall, specificity, weighted F1-score, melanoma-specific precision/recall/F1, and ROC-AUC. Recall for the melanoma class is equivalent to sensitivity in this binary formulation. In addition to classification quality, training time and average per-image inference time were recorded to assess deployment feasibility.

3.6 Hyperparameter Configuration

The principal hyperparameters and implementation settings used in the final experiments are summarized in Table 1. All values in this table were taken directly from

the Python notebook used for model training and evaluation.

Table -1: Hyperparameter settings used in model training and evaluation

Parameter	Value / Setting
Random seed	42
Input image size	224 × 224 × 3
Training/validation batch size	32
Test batch size	1
Phase 1 training	15 epochs with frozen backbone
Phase 2 fine-tuning	35 epochs with top 30% of backbone layers unfrozen; BatchNorm layers kept frozen
Optimizer	Adam
Head learning rate	1e-3
Fine-tuning learning rate	1e-5
Loss function	Binary cross-entropy (label smoothing = 0.0)
Class imbalance handling	Balanced class weights enabled
Training augmentation	Rotation = 20, width shift = 0.03, height shift = 0.03, zoom = 0.08, horizontal flip
Threshold search	Thresholds from 0.10 to 0.90 in 161 steps; minimum validation specificity = 0.95
Test-time augmentation	5 passes total: 1 original + 4 augmented passes with horizontal flip, rotation = 10, zoom = 0.05

4. RESULTS AND DISCUSSION

The experimental results are illustrated in Figures 1-10, summarized numerically in Table 2 and Table 3, and finally synthesized in Figure 11. EfficientNetB3 achieved the highest overall accuracy, weighted F1-score,

specificity, and ROC-AUC, whereas MobileNetV2_S remained competitive while requiring fewer parameters and lower inference time.

This indicates a practical trade-off for clinical deployment. EfficientNetB3 is the stronger overall model when maximum predictive performance and balanced accuracy are required. Conversely, MobileNetV2_S achieved a slightly higher melanoma recall (93.70% vs. 92.84%), making it a highly sensitive, lightweight option suited for first-line screening environments where minimizing false negatives (missed melanomas) is the highest clinical priority.

Furthermore, the proposed EfficientNetB3 pipeline demonstrated competitive performance relative to recent literature. Kaur et al. [1] reported 93.40% classification accuracy in their melanoma CAD framework, while Mohankumar et al. [2] reported 92% validation accuracy in a deep learning system using VGG19 and MobileNetV2. In comparison, the proposed EfficientNetB3 model achieved 94.72% test accuracy and 0.9860 ROC-AUC on ISIC2020. Although direct comparison across studies should be interpreted carefully because of differences in preprocessing, splits, and evaluation settings, these results indicate that the proposed combination of class weighting, two-stage fine-tuning, threshold optimization, and five-pass test-time augmentation provides a robust framework for melanoma classification.

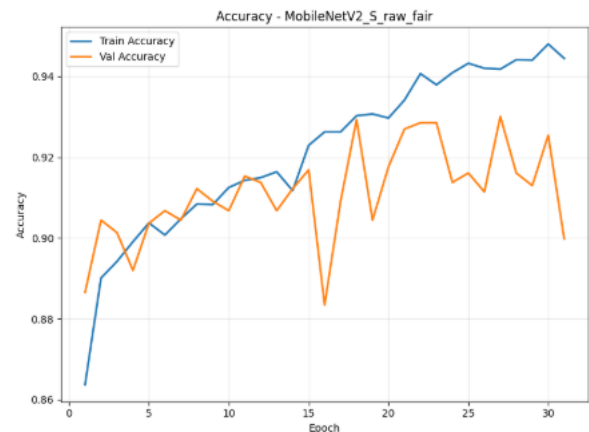


Fig -1: MobileNetV2_S accuracy history from the Python notebook output.

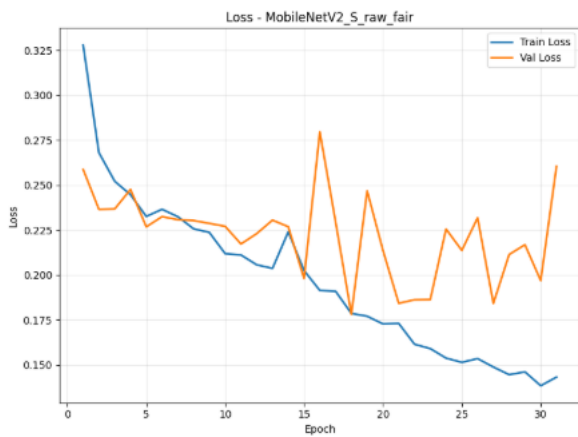


Fig -2: MobileNetV2_S loss history from the Python notebook output.

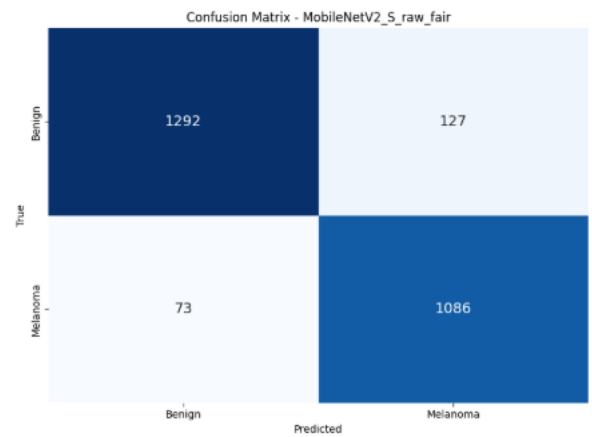


Fig -5: MobileNetV2_S confusion matrix from the Python notebook output.

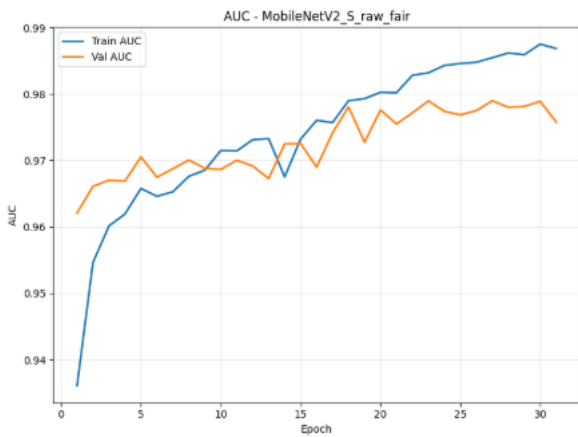


Fig -3: MobileNetV2_S AUC history from the Python notebook output.

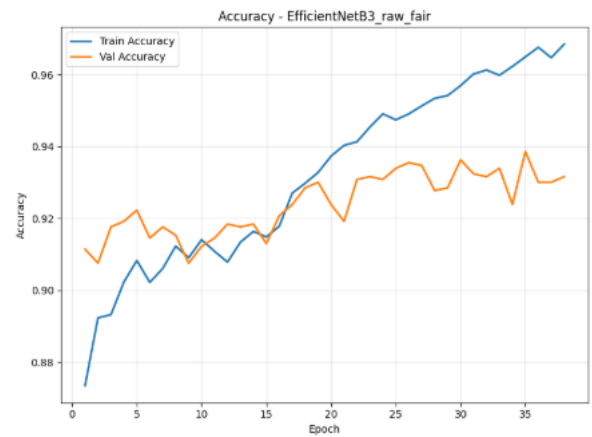


Fig -6: EfficientNetB3 accuracy history from the Python notebook output.

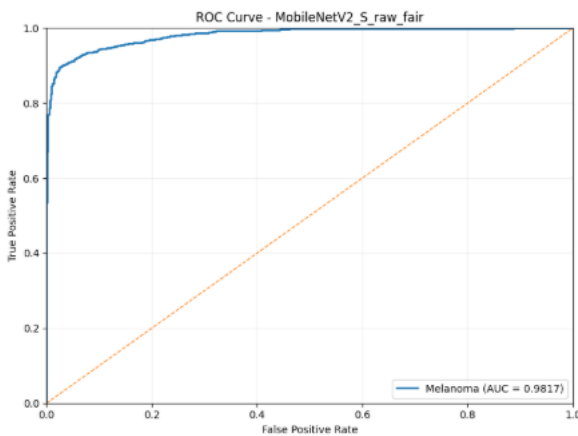


Fig -4: MobileNetV2_S ROC curve from the Python notebook output.

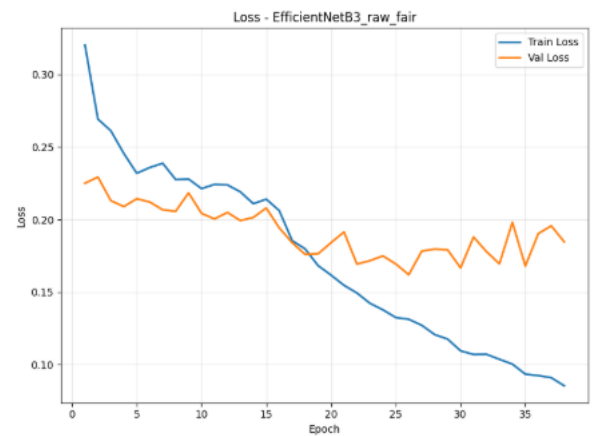


Fig -7: EfficientNetB3 loss history from the Python notebook output.

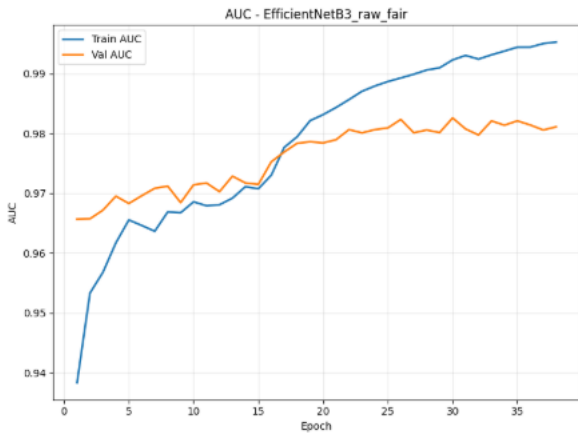


Fig -8: EfficientNetB3 AUC history from the Python notebook output.

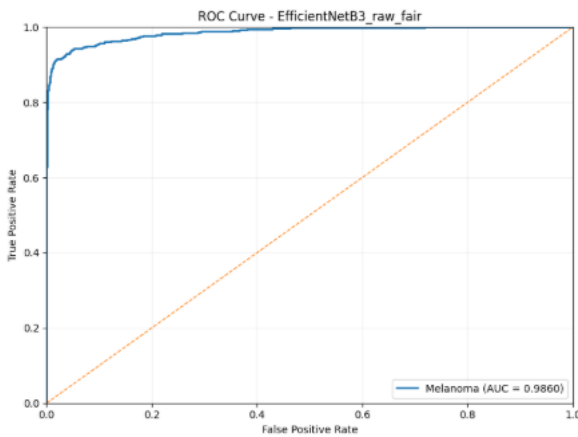


Fig -9: EfficientNetB3 ROC curve from the Python notebook output.

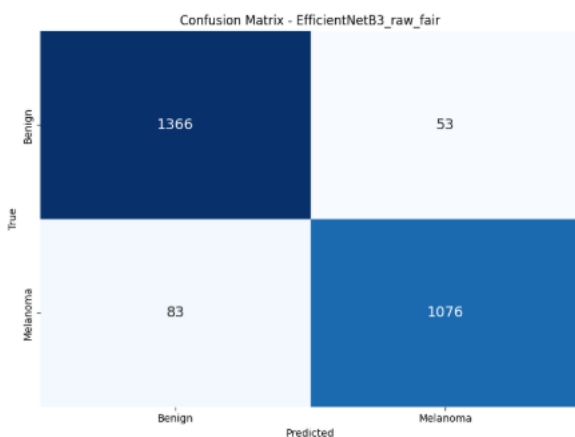


Fig -10: EfficientNetB3 confusion matrix from the Python notebook output.

Table -2: Comparative Test Performance of the Proposed Models

Metric	MobileNetV2_S	EfficientNetB3
Threshold (Thr.)	0.350	0.495
Accuracy (ACC)	92.24	94.72
Precision (PRE)	92.35	94.74
Recall (REC)	92.24	94.72
Specificity (SPE)	91.05	96.26
F1-Score (F1)	92.26	94.72
AUC	0.9817	0.9860

Table -3: Melanoma-Specific and Computational Performance of the Proposed Models

Metric	MobileNetV2_S	EfficientNetB3
Melanoma Precision (M-PRE)	89.53	95.31
Melanoma Recall (M-REC)	93.70	92.84
Melanoma F1-Score (M-F1)	91.57	94.06
Parameters (Millions)	2.18	11.71
Training Time (Mins)	59.43	80.21
Test Time per Image (Sec)	0.0805	0.1125

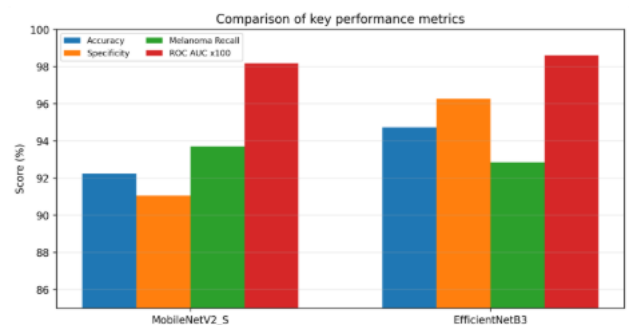


Fig -11: Comparison of key performance metrics derived from the project outputs.

MobileNetV2_S reached 92.24% test accuracy, 91.05% specificity, and 93.70% melanoma recall at an optimized threshold of 0.350. Its confusion matrix shows 1,292 benign lesions correctly identified, 127 benign lesions misclassified as melanoma, 1,086 melanoma lesions correctly identified, and 73 melanoma lesions missed. The best validation accuracy observed during training was approximately 93.01%, with a best validation AUC of 97.90%.

EfficientNetB3 produced the best overall test result with 94.72% accuracy, 96.26% specificity, 95.31% melanoma precision, and 0.9860 ROC-AUC at a threshold of 0.495. Its confusion matrix shows 1,366 benign lesions correctly classified, only 53 benign false alarms, 1,076 melanoma lesions correctly classified, and 83 melanoma misses. The model also achieved a higher best validation accuracy of about 93.86% and a best validation AUC of 98.26%.

The training curves further show that both models converged smoothly after transfer learning, with gradual improvement in training accuracy and AUC. EfficientNetB3 maintained a stronger validation trend and lower false-positive burden, which explains its superior specificity. However, MobileNetV2_S recorded a slightly higher melanoma recall, meaning it missed fewer melanoma cases on the held-out test set. This distinction is important in clinical settings: depending on whether the application prioritizes sensitivity or overall balance, either model may be preferred. The slightly higher final test AUC values relative to the best validation of AUC are plausible because final testing used five-pass test-time augmentation, which stabilized prediction scores.

From a computational perspective, MobileNetV2_S completed training in about 59.43 minutes and required 0.0805 seconds per test image, whereas EfficientNetB3 required 80.21 minutes and 0.1125 seconds per image. The additional cost of EfficientNetB3 is acceptable for workstation deployment, but MobileNetV2_S remains advantageous when memory and latency constraints are strict.

5. CONCLUSIONS

This work presented a study on skin cancer detection using deep learning with two transfer-learning classifiers, MobileNetV2_S and EfficientNetB3, trained on ISIC2020 dermoscopic images. The implemented pipeline combines data augmentation, class balancing, staged fine-tuning, validation-based threshold selection, and test-time augmentation to improve robustness.

Among the evaluated models, EfficientNetB3 delivered the best overall classification performance with 94.72% accuracy, 96.26% specificity, and 0.9860 ROC-AUC, while MobileNetV2_S offered a lighter alternative with lower computational demand and slightly higher melanoma recall. These results show that transfer learning can provide practical support for fast and consistent

melanoma screening from dermoscopic images. Despite these promising results, the present evaluation is limited to binary benign-versus-melanoma classification on a single public dataset and should be complemented by external clinical validation before deployment.

Future work can extend the present system by incorporating lesion segmentation, explainability methods such as Grad-CAM, external clinical validation, and multi-class skin lesion classification. Integration into a secure clinical interface or mobile-assisted workflow may further improve accessibility in underserved regions.

ACKNOWLEDGEMENT

The authors acknowledge the ISIC archive for providing access to dermoscopic image data and the open-source TensorFlow ecosystem used for model development and evaluation.

REFERENCES

- [1] R. Kaur, H. GholamHosseini, and M. Linden, "Advanced Deep Learning Models for Melanoma Diagnosis in Computer-Aided Skin Cancer Detection," *Sensors*, vol. 25, no. 3, Art. no. 594, 2025.
- [2] L. Mohankumar, K. Lakshmi Saraswathi, and K. V. Ramana, "Skin Cancer Detection by Using Deep Learning," *International Journal of Innovative Research in Technology*, vol. 11, no. 11, pp. 4775-4779, 2025.
- [3] V. Rotemberg et al., "A patient-centric dataset of images and metadata for identifying melanomas using clinical context," *Scientific Data*, vol. 8, Art. no. 34, 2021.
- [4] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted Residuals and Linear Bottlenecks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 4510-4520.
- [5] M. Tan and Q. V. Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks," in *Proc. 36th Int. Conf. Mach. Learn. (ICML)*, PMLR, vol. 97, 2019, pp. 6105-6114.
- [6] M. Abadi et al., "TensorFlow: A System for Large-Scale Machine Learning," in *Proc. 12th USENIX Symp. Operating Systems Design and Implementation (OSDI)*, 2016, pp. 265-283.
- [7] B. Cassidy, C. Kendrick, A. Brodzicki, J. Jaworek-Korjakowska, and M. H. Yap, "Analysis of the ISIC image datasets: Usage, benchmarks and recommendations," *Medical Image Analysis*, vol. 75, Art. no. 102305, 2022.