

Improving Dermoscopic Melanoma Recall via a Weighted Xception–EfficientNetB4 Ensemble and Test-Time Augmentation

Abdulmunem A. Alkeelani¹, Abdulgani Albagul^{1,2*}

¹College of Electronic Technology Baniwalid, Baniwalid, Libya

²Libyan Authority for Scientific Research, Tripoli, Libya

Email: ¹a.alkeelani@almadar.ly, ²albagoul@yahoo.com

Received: June 14, 2026

Accepted: September 25, 2026

Published: October 20, 2026

DOI: 10.65998/ijees.v4i2.194

Abstract: Automated dermoscopic image classification can support skin-lesion assessment, but severe class imbalance in public datasets can conceal poor melanoma sensitivity behind high overall accuracy. This study investigated whether combining complementary transfer-learning architectures with class-aware augmentation and test-time augmentation (TTA) could improve melanoma detection. A four-class subset of the ISIC 2019 archive containing 23,346 images of melanoma (MEL), melanocytic nevus (NV), basal cell carcinoma (BCC), and benign keratosis (BKL) was evaluated. Images underwent automated hair removal, contrast enhancement, colour normalization, and architecture-specific resizing. A custom residual CNN, EfficientNetB4, Xception, and a weighted Xception–EfficientNetB4 ensemble were compared on the same held-out test set. Performance was assessed using accuracy, precision, recall, F1-score, macro-AUC, Wilson confidence intervals, DeLong's test, McNemar's test, and expected calibration error (ECE). The ensemble with TTA achieved 89.45% accuracy, 88.36% macro-F1, 88.62% weighted-F1, and 0.9655 macro-AUC. Melanoma recall increased from 70.71% with the custom CNN to 84.57% with the final ensemble. Melanoma misclassification as nevus decreased from 162 to 80 cases. The ensemble also had the lowest reported ECE (0.08). Weighted ensembling with TTA improved classification performance and melanoma sensitivity under the evaluated conditions. The residual melanoma miss rate and absence of external clinical validation indicate that the system should be considered a decision-support research model rather than an autonomous diagnostic system.

Keywords: Melanoma detection; dermoscopy; deep learning; transfer learning; ensemble learning.

1. Introduction

Skin cancer remains among the most frequently diagnosed malignancies worldwide. Melanoma, although accounting for only about 1% of all skin cancer cases, is responsible for the substantial majority of skin cancer-related deaths. In the United States alone, an estimated 104,960 new cases of invasive melanoma were projected for 2025, with approximately 8,430 associated deaths [1]. The clinical stakes of early detection are severe: five-year survival exceeds 99% for localised disease but falls to approximately 35% once the disease has metastasised. This asymmetry between the curability of early-stage disease and the lethality of late-stage disease is the central motivation for improving diagnostic sensitivity at the point of first assessment. Moreover, diagnosis conventionally proceeds through visual examination, often supported by dermoscopy, and is constrained by well-documented limits on human sensitivity. A large systematic review and meta-analysis reported a melanoma-detection sensitivity of only 85.7% for experienced dermatologists and just 37.5% for primary-care physicians [2]. Excisional

biopsy with histopathological review remains the diagnostic gold standard, but it is invasive, costly, and frequently inaccessible in under-resourced settings. These constraints motivate automated, image-based decision-support tools capable of flagging suspicious lesions for priority review.

Convolutional neural networks (CNNs) have repeatedly demonstrated dermatologist-level discriminative performance on curated dermoscopic benchmarks [3,4]. However, a critical and frequently underreported limitation of this literature is that headline accuracy figures are dominated by the majority benign class. Public archives such as ISIC 2019 exhibit an approximate 3:1 ratio of benign nevi to melanoma [5-7], and models optimised for aggregate accuracy on such data systematically under-detect the minority malignant class, the opposite of what is clinically required. Addressing this sensitivity gap, rather than further inflating overall accuracy, is the specific problem this study addresses.

Transfer learning from ImageNet-pretrained backbones [8-12], ensemble methods [13], and test-time augmentation are widely used to improve dermoscopic classification, but their combined effect on melanoma recall under class imbalance has received little attention; the relevant literature is reviewed in Section 2. That review identifies five gaps (Section 2.4): few studies target melanoma recall as the primary criterion [5]; weighted ensembling combined with test-time augmentation is rarely examined for melanoma; technical metrics are seldom translated into clinically interpretable terms [2]; cross-study comparisons are confounded by differing protocols [2]; and calibration is rarely assessed alongside discrimination [14-17]. The present study addresses these gaps within a single, internally consistent experimental framework.

Accordingly, three research questions are posed. RQ1: to what extent does class imbalance reduce melanoma recall relative to overall accuracy for a custom CNN trained on the ISIC 2019 four-class subset? RQ2: does a weighted ensemble of Xception and EfficientNetB4 improve melanoma recall beyond either constituent model? RQ3: does test-time augmentation further improve classification performance and melanoma sensitivity?

2. Related Work

This section reviews the clinical context of melanoma diagnosis, the public dermoscopic datasets and their class-imbalance problem, the deep-learning architectures used for skin-lesion classification, and the research gaps that motivate the present study.

2.1 Clinical Diagnostic Accuracy and the Case for Computer-Aided Diagnosis

Melanoma arises from epidermal melanocytes and, owing to its aggressive metastatic potential to lung, liver, brain, and bone, disproportionately drives skin-cancer mortality despite its low relative incidence [1]. Basal cell carcinoma (BCC) is the most common skin malignancy and is locally invasive but rarely metastatic, whereas squamous cell carcinoma exhibits intermediate metastatic risk. Dermoscopy, which uses polarised or immersion-contact magnification (typically 10×–100×) to visualise subsurface structures invisible to the unaided eye, measurably improves diagnostic accuracy: dermoscopy-assisted examination by experienced dermatologists achieves 85.7% sensitivity and 81.3% specificity for melanoma, compared with markedly lower performance for unaided clinical examination, particularly among less experienced examiners and primary-care physicians (37.5% sensitivity) [2], [7]. Structured algorithms, the ABCDE rule, the 7-point checklist, the Menzies method, and the chaos-and-clues approach, improve consistency but do not eliminate inter-observer variability. These residual gaps in human diagnostic sensitivity, combined with the invasiveness and limited accessibility of confirmatory biopsy, constitute the clinical rationale for automated, image-based decision support.

2.2 Public Dermoscopic Datasets and the Class-Imbalance Problem

Progress in this field has been driven substantially by publicly released, ISIC-curated benchmark datasets of increasing scale as demonstrated in Table 1. The ISIC 2019 challenge dataset, aggregating imagery from the HAM10000 collection, the BCN20000 archive of Hospital Clinic de Barcelona, and the MSKCC dataset, comprises 25,331 dermoscopic images across eight diagnostic categories, of which the

four used in the present study total 23,346 images [5,18-20]. A limitation common to every major public archive is severe class imbalance: melanocytic nevi outnumber melanoma cases by ratios ranging from approximately 3:1 (ISIC 2019, the four-category subset used here) to 6:1 (HAM10000) [5]. Because standard training objectives optimise aggregate accuracy, this imbalance biases learned decision boundaries toward the majority benign class, producing models that report impressive overall accuracy (frequently exceeding 95% on curated benchmarks) while masking clinically unacceptable melanoma recall [6]. A second, related limitation is the underrepresentation of darker Fitzpatrick skin types across all major archives, which constrains generalisability and raises equity concerns for real-world deployment.

Table 1. Major publicly available dermoscopic image datasets.

Dataset	Year	Images	Melanoma Cases	Diagnostic Classes
PH2	2013	200	40	3
ISIC 2016	2016	900	273	2
ISIC 2017	2017	2000	374	3
HAM10000	2018	10015	1113	7
ISIC 2019	2019	25331	4522	8
ISIC 2020	2020	33126	584	2
ISIC2024	2024	401059	>10000	9

2.3 Deep Learning Architectures for Skin Lesion Classification

The seminal study of Esteva et al. Demonstrated that a CNN trained on 129,450 clinical images could match the performance of 21 board-certified dermatologists (AUC 0.91) on benign-versus-malignant discrimination [4]. Because annotated medical imagery is comparatively scarce, transfer learning from ImageNet-pretrained backbones has become the dominant paradigm: low-level features (edges, textures, colour gradients) generalise well from natural images, while higher-level representations are adapted during fine-tuning [8]. Widely benchmarked backbones include VGG (uniform stacked 3×3 convolutions) [9], ResNet (residual skip connections) [10], DenseNet (dense feature reuse) [11], and EfficientNet, which applies a compound scaling coefficient across network depth, width, and input resolution [12]. Xception replaces standard convolutions with depthwise-separable convolutions, improving parameter efficiency while preserving fine-grained spatial discrimination [21-29].

Ensemble methods, which combine the predictions of architecturally diverse models, offer complementary error profiles and have shown improved robustness in several domains [13]. Hybrid architectures, CNN-LSTM combinations for sequential feature modelling [14], and attention mechanisms such as Squeeze-and-Excitation blocks [15] represent further directions. Vision Transformers (ViTs), which model global image context via self-attention over image patches, have shown competitive or superior performance to CNNs on some histopathological tasks but typically require substantially larger training corpora [16]. Across this literature, two concerns recur: the ‘black-box’ opacity of deep models, which limits clinician trust and complicates regulatory approval absent explainability tooling such as Grad-CAM, and probabilistic miscalibration, whereby predicted confidence diverges from true outcome frequency in ways that can mislead clinical users [17].

2.4 Research Gaps

Synthesising this literature identifies five gaps that motivate the present study. First, despite widespread reporting of high aggregate accuracy, few studies explicitly target melanoma recall as the primary optimisation and evaluation criterion, even though it is the metric of greatest clinical consequence [5]. Second, systematic optimisation of ensemble weighting combined with Test-Time Augmentation remains rare specifically for melanoma detection. Third, technical metrics are seldom translated into clinically interpretable terms, such as the number of additional true-positive melanomas identified per fixed case volume [2]. Fourth, cross-study comparisons are frequently confounded by differing datasets, class definitions, and evaluation protocols, impeding fair architectural comparison

[2]; controlled, same-dataset comparison of custom, transfer-learned, and ensemble architectures is uncommon. Fifth, model calibration, essential for trustworthy probabilistic decision support, is rarely assessed alongside discriminative performance [17]. The present study is designed to address all five gaps within a single, internally consistent experimental framework.

3. Materials and Methods

This section describes the materials, methodological framework, and experimental procedures used to develop and evaluate the proposed skin lesion classification approach. It presents the dataset and its composition, image preprocessing and augmentation procedures, class-imbalance mitigation strategy, model architectures, ensemble configuration, and test-time augmentation (TTA) procedure. The training protocol and evaluation metrics are also detailed to provide a clear and reproducible account of the experimental design.

3.1 Study Design

This was a retrospective computational study based on publicly available dermoscopic images. The analysis consisted of model development, internal validation, and evaluation on a held-out test subset. The study should be understood as an internal technical evaluation rather than a prospective clinical validation study. The research follows a quantitative experimental methodology comprising five sequential phases: data acquisition and pre-processing, data augmentation, model development, training and optimization, and final evaluation as shown in Figure 1.

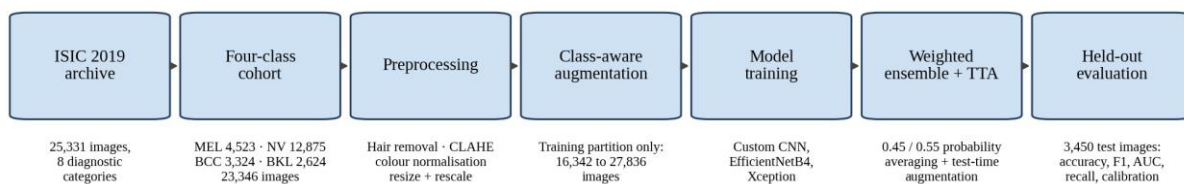


Figure 1. Overall workflow of the study.

3.2 Dataset

This study uses the ISIC 2019 Challenge dataset, restricted to four diagnostically important and visually confusable categories: melanoma (MEL, malignant), melanocytic nevus (NV, benign), basal cell carcinoma (BCC, malignant), and benign keratosis (BKL, benign; encompassing seborrheic keratosis and solar lentigo) [18,19]. The resulting four-class subset comprises 23,346 dermoscopic images sourced from three complementary acquisition programmes—the HAM10000 collection (Medical University of Vienna and an Australian skin-cancer clinic), the BCN20000 archive (Hospital Clinic de Barcelona), and the MSKCC dataset, captured with standardised dermoscopy systems (FotoFinder, DermLite DL3, Canon PowerShot G16 with DermLite attachment, Heine Delta 20T) at 10×–100× magnification [5,19,20]. Diagnostic ground truth was established by histopathological confirmation, reflectance confocal microscopy with expert interpretation, or documented multi-expert consensus for unequivocally benign lesions. All images are de-identified and distributed under Creative Commons licensing for academic use, with institutional approvals and informed consent obtained by the contributing centres. The dataset was partitioned by stratified random sampling (fixed seed = 42) into training (70%), validation (15%), and held-out test (15%) subsets, preserving the native class distribution in the validation and test partitions to ensure an evaluation representative of real-world prevalence; patient-level separation across partitions was enforced to prevent leakage. The class-wise composition and the final test-set sizes actually used for model evaluation are reported in Table 2.

Table 2. ISIC 2019 four-class subset composition and held-out test-set sizes used for final evaluation.

Diagnostic Class	Type	Original Count	Test-Set Size (Evaluation)
Melanoma (MEL)	Malignant	4,523	700
Melanocytic Nevus (NV)	Benign	12,875	1,843
Basal Cell Carcinoma (BCC)	Malignant	3,324	515
Benign Keratosis (BKL)	Benign	2,624	392
Total	—	23,346	3,450

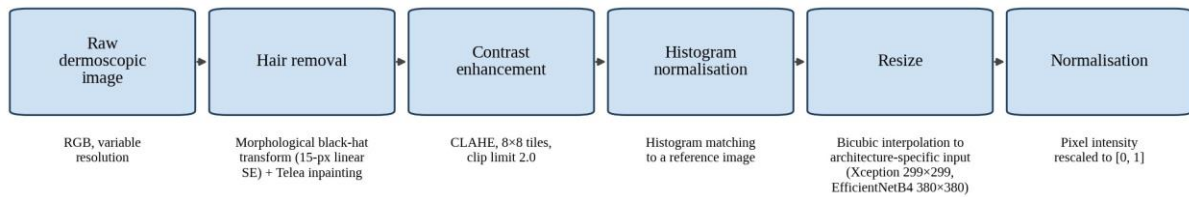
Note: the nominal 70/15/15 split targets 3,502 validation and 3,502 test images (16,342 training); the 3,450-image test set reported here reflects the final evaluation partition after preprocessing-stage exclusions and is the basis for all results in Section 4.

3.3 Image Preprocessing

Dermoscopic images present acquisition-specific challenges, hair fibres, air bubbles, ruler markings, specular reflections, vignetting, and inter-device colour and contrast variation, that can obscure diagnostically relevant structures such as pigment networks and regression areas. A sequential, fully automated preprocessing pipeline was applied uniformly to all images as illustrated in Figure 2: Hair removal: a morphological black-hat transform $BlackHat(I) = (I \oplus SE) \ominus SE - I$, where $\oplus$ and $\ominus$ denote dilation and erosion and SE is a 15-pixel linear structuring element) isolates hair structures; the resulting mask is thresholded, mildly dilated, and used to guide Telea inpainting, which reconstructs the masked regions from surrounding pixel information [21].

Contrast enhancement: Contrast-Limited Adaptive Histogram Equalisation (CLAHE) is applied per RGB channel using 8×8-pixel tiles and a clip limit of 2.0, selected empirically to enhance local contrast without amplifying sensor noise [22]. Colour normalisation: histogram matching aligns each image's cumulative distribution function (CDF) to a reference image selected as the archive item with minimum mean histogram distance to all others; for source intensity v_s and reference CDF F_r , the mapped intensity v_m satisfies $F_r(v_m) = F_s(v_s)$, which preserves within-channel intensity ordering while harmonising inter-device colour variation.

Resizing and normalisation: bicubic interpolation resizes each image to the architecture-specific input resolution (299×299 for Xception and 380×380 for EfficientNetB4), after which pixel intensities are linearly rescaled to [0, 1].

**Figure 2.** Image preprocessing pipeline: hair removal.

3.4 Data Augmentation and Class-Imbalance Mitigation

Figure 3 shows the class distribution of the four diagnostic categories, highlighting the severe imbalance (NV: 12,875 vs. BKL: 2,624) that motivates the class-aware augmentation strategy. The melanocytic nevus (NV) class substantially outnumbers the melanoma (MEL), basal cell carcinoma (BCC), and benign keratosis (BKL) classes, creating an approximately 3:1 nevus-to-melanoma imbalance that biases standard training objectives toward the majority class.

Online (on-the-fly) augmentation was implemented via Keras/TensorFlow data generators [23], avoiding storage overhead and ensuring every training epoch presents novel augmented variants. Geometric transformations comprised random rotation ($\pm 180^\circ$), horizontal/vertical flipping ($p = 0.5$ per axis), scaling ($0.8 \times$ – $1.2 \times$), and translation (up to 10% of image extent). Photometric transformations comprised brightness, contrast, and saturation jitter (each sampled uniformly from [0.8, 1.2]), hue offset (± 0.1), and additive Gaussian noise ($\sigma = 0.01$). Two higher-order techniques were additionally applied: CutOut [30], which occludes random square regions to discourage over-reliance on any single

discriminative subregion, and MixUp, which forms convex combinations of paired training images and their one-hot labels, $x' = \lambda x_i + (1 - \lambda)x_j$ and $y' = \lambda y_i + (1 - \lambda)y_j$, with $\lambda \sim \text{Beta}(\alpha, \alpha)$ and $\alpha = 0.2$ [24].

To counteract the native 3:1 (NV:MEL) and greater imbalance across classes, augmentation intensity was scaled inversely to class frequency (class-aware augmentation), applying stronger and more frequent transformations to BCC and BKL (3× effective multiplier) and MEL (2×) than to the majority NV class (1×), yielding a substantially rebalanced effective training set of 27,836 images while the validation and test partitions retained the natural, unaugmented class distribution to preserve a realistic evaluation.

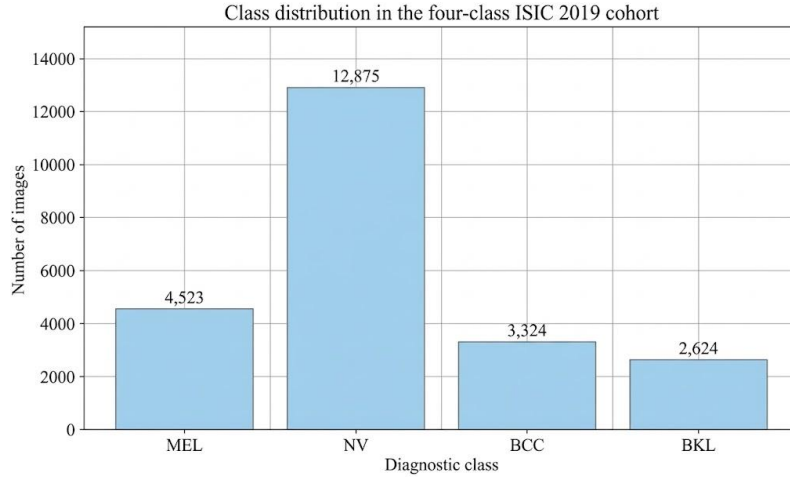


Figure 3. Class distribution of the four diagnostic categories used in the study.

3.5 Model Architectures

A custom convolutional architecture (Advanced_Balanced_Model, $\approx 14.3\text{M}$ parameters) with residual connections and dual (average/max) pooling was trained from random initialisation as a non-transfer-learning baseline, sized to be parameter-comparable with the pretrained models so that performance differences could be attributed to pretraining rather than capacity. Figure 4 summarises the full architecture and ensemble strategy described below.

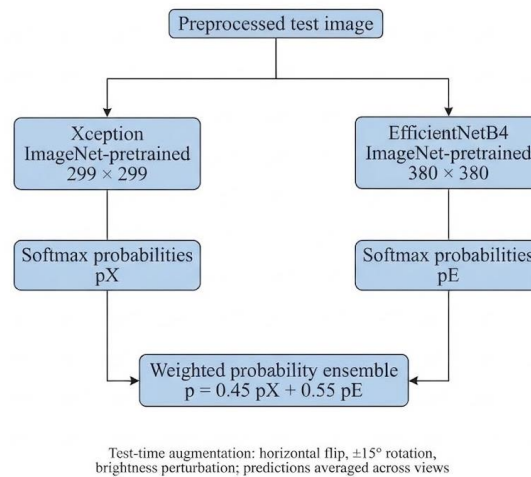


Figure 4. Weighted Xception-EfficientNetB4 ensemble and test-time augmentation strategy.

EfficientNetB4 ($\approx 18.6\text{M}$ parameters) and Xception ($\approx 22\text{M}$ parameters) were both initialised from ImageNet pre-training and fine-tuned under identical mixed-precision training to avoid uncontrolled training differences. EfficientNetB4 applies compound scaling across depth, width, and input resolution [12]; Xception replaces standard convolutions with depthwise-separable convolutions, decoupling spatial and cross-channel feature learning to improve parameter efficiency while retaining sensitivity

to fine-grained spatial structure [29]. The final model is a weighted average of the two networks' softmax outputs:

$$p_{ensemble} = 0.45 \cdot p_{Xception} + 0.55 \cdot p_{EfficientNetB4}$$

With the larger weight assigned to EfficientNetB4, which achieved the higher standalone validation accuracy. Weights were selected by grid search on the validation set under a sum-to-one constraint; all ensemble and augmentation decisions used validation data only, with the test set reserved exclusively for final evaluation.

At inference, each pre-processed test image is passed through Xception (299×299 input) and EfficientNetB4 (380×380 input). Test-Time Augmentation (TTA) averages predictions across horizontal flips, fixed rotations of -15° , 0° , and $+15^\circ$, and brightness perturbations over 0.8–1.2, reducing prediction variance on borderline cases. This pairing was chosen for complementarity rather than raw capacity: Xception's depthwise-separable convolutions emphasise fine-grained local texture, while EfficientNetB4's compound scaling captures multi-scale contextual structure.

3.6 Training Configuration

All models were trained with the Adam optimiser and categorical cross-entropy loss. Learning-rate scheduling used ReduceLROnPlateau, and early stopping (monitored on validation loss) was applied to control overfitting; the custom CNN was trained for up to 100 epochs, EfficientNetB4 for up to 50 epochs (best checkpoint at epoch 25: 87.0% validation accuracy, 0.965 validation AUC), and Xception for up to 30 epochs (best validation accuracy 84.9% at epoch 29, validation AUC 0.961). Training was conducted on a CUDA-enabled NVIDIA GPU using TensorFlow 2.x with the Keras API; mixed-precision computation was enabled for EfficientNetB4 to improve training throughput.

3.7 Evaluation Metrics and Statistical Analysis

Discriminative performance was quantified using overall accuracy, per-class precision, recall (sensitivity), F1-score, macro- and weighted-average aggregates [25], and AUC-ROC. Because melanoma recall is the clinically decisive metric, it is reported and discussed independently of aggregate accuracy throughout. Three complementary statistical procedures were used to test whether observed differences between models reflect genuine performance gaps rather than sampling variation: (i) 95% confidence intervals for overall accuracy, computed via the Wilson score interval; (ii) DeLong's non-parametric test for comparing correlated AUC estimates [26]; and (iii) McNemar's test for paired binary-outcome comparison on the same test set [27].

Probabilistic reliability was assessed via the Expected Calibration Error (ECE), which quantifies the mean absolute discrepancy between predicted confidence and empirical outcome frequency across confidence bins, following Guo et al. [28]; a well-calibrated model assigning 0.8 confidence to a melanoma prediction should be correct in approximately 80% of such instances. Finally, clinical impact was expressed directly as the estimated number of additional true-positive melanoma detections per 700 evaluated cases, obtained from the difference in melanoma recall between the ensemble and the non-transfer-learning baseline.

4. Results

This section presents the experimental results obtained from the evaluated classification models on the held-out test set. The analysis begins with the composition of the evaluation cohort and overall model performance, followed by class-specific performance, melanoma sensitivity, confusion-matrix analysis, and the contribution of TTA. Statistical comparisons, calibration results, and the potential clinical implications of the observed classification errors are also reported.

4.1 Experimental Setup

All four architectures, the custom CNN, EfficientNetB4, Xception, and the weighted Xception–EfficientNetB4 ensemble (with and without TTA), were evaluated on the identical held-out test set of

3,450 images (700 MEL, 1,843 NV, 515 BCC, 392 BKL; [Table 2](#)), using the metrics and statistical procedures described in Section 3.7.

4.2 Individual Model Performance

The custom CNN, trained from random initialisation, converged to 98% training accuracy with validation accuracy plateauing near 95% after approximately 60 of its 100 training epochs, although this level was not reproduced on the held-out test set (see Section 5.4), where it achieved only 84.49% overall accuracy and, critically, 70.71% melanoma recall, illustrating directly how class imbalance suppresses minority-class sensitivity even when aggregate training behaviour appears healthy. EfficientNetB4 reached 94.1% training accuracy with validation accuracy peaking at 87.0% at epoch 25 before a widening train–validation gap indicated the onset of mild overfitting; on the test set it achieved 87.05% overall accuracy, 96.49% macro-AUC, and 82.86% melanoma recall. Xception achieved 90.8% training accuracy and 84.9% peak validation accuracy (epoch 29, validation AUC 0.961), translating to 85.68% test accuracy, 96.32% macro-AUC, and 82.00% melanoma recall. Both pretrained backbones outperformed the custom CNN by a clear margin, empirically confirming the value of ImageNet transfer learning for this domain: EfficientNetB4 exceeded the custom baseline by 2.56 percentage points in overall accuracy despite the latter’s larger effective training budget.

4.3 Ensemble Performance

The weighted ensemble without TTA already surpassed every individual model, reaching 88.70% overall accuracy, 86.87% macro-F1, and 87.71% weighted-F1. Applying TTA yielded a further, consistent improvement across all aggregate metrics: overall accuracy rose to 89.45% (+0.75 points), macro-F1 to 88.36% (+1.49 points), and weighted-F1 to 88.62% (+0.91 points) as highlighted in [Table 3](#) and [Figure 5](#). Per-class performance for the final ensemble-with-TTA configuration is reported in [Table 4](#), and the corresponding confusion matrix is shown in [Figure 6](#). Melanoma recall for each configuration, together with published clinical benchmarks, is compared in [Figure 7](#).

Table 3. Overall performance comparison across all evaluated model configurations on the held-out test set

Model	Overall Accuracy	Macro-Avg F1	Weighted-Avg F1	Macro-Avg AUC
Custom CNN (baseline)	84.49%	81.35%	84.39%	0.9533
EfficientNetB4	87.05%	85.65%	86.11%	0.9649
Xception	85.68%	83.49%	85.92%	0.9632
Ensemble (no TTA)	88.70%	86.87%	87.71%	
Ensemble + TTA (final)	89.45%	88.36%	88.62%	0.9655

The ensemble with TTA achieves the highest overall accuracy (89.45%), macro-F1 (88.36%), and macro-AUC (96.55%) compared to all individual models. The progressive improvement from the custom CNN baseline through the transfer-learned models to the final ensemble demonstrates the value of both transfer learning and ensembling.

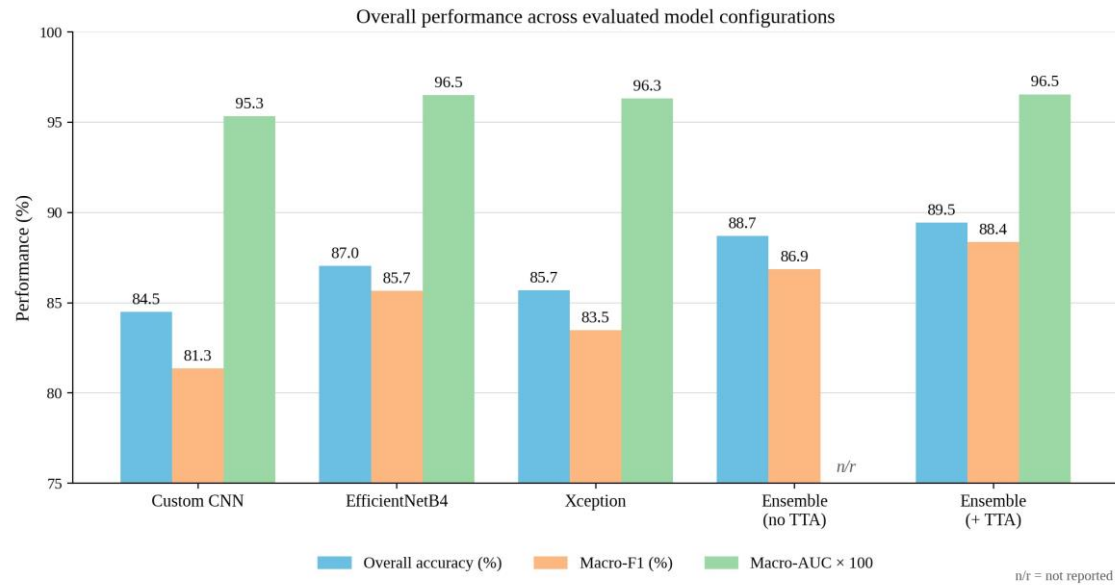


Figure 5. Overall accuracy, macro-F1 and macro-AUC across the evaluated model configurations.

Table 4. Per-class performance of the final ensemble model (weighted Xception-EfficientNetB4 + TTA).

Class	Precision	Recall (Sensitivity)	F1-Score	Support
MEL (melanoma)	0.8385	0.8457	0.8421	700
NV (nevus)	0.8953	0.9143	0.9047	1,843
BCC	0.9326	0.9670	0.9495	515
BKL	0.8066	0.8724	0.8382	392
Macro average	0.8683	0.8999	0.8836	3,450
Weighted average	0.8790	0.8945	0.8862	3,450

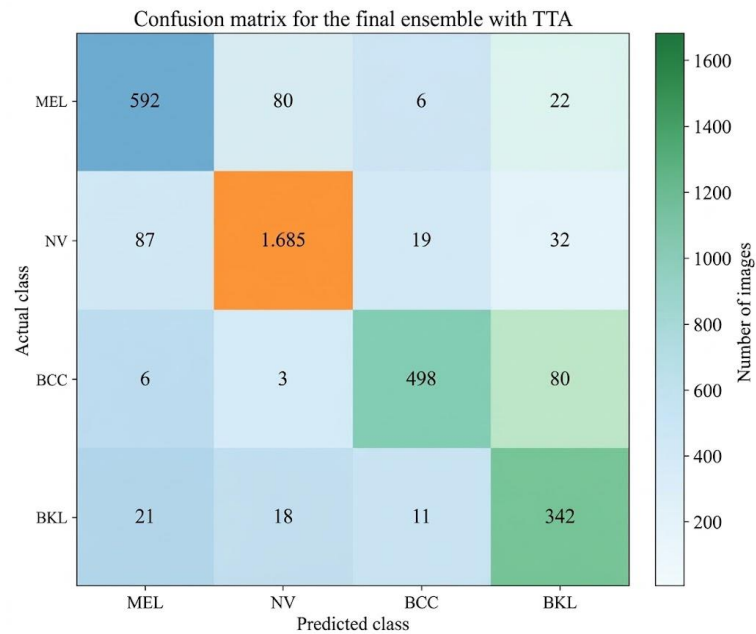


Figure 6. Confusion matrix for the final weighted Xception-EfficientNetB4 ensemble with TTA.

The model correctly classified 592 of 700 melanoma cases (84.57% recall), with 80 melanoma cases misclassified as nevus (11.4%), a substantial reduction from the baseline's 162 cases (23.1%). BCC detection was near-perfect (96.70% recall).

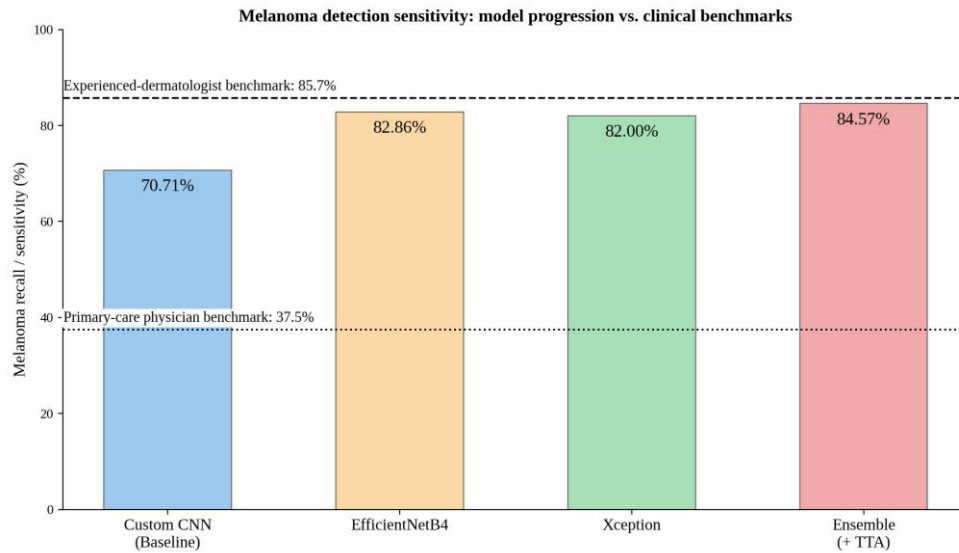


Figure 7. Melanoma recall (sensitivity) across the evaluated model configurations.

The ensemble with TTA achieves 84.57% melanoma recall, approaching the 85.7% benchmark for experienced dermatologists and substantially exceeding the 37.5% sensitivity of primary-care physicians.

4.4 Statistical Significance Testing

Wilson-score 95% confidence intervals (Table 5) show non-overlapping intervals between the ensemble and every individual model, consistent with a genuine accuracy improvement rather than sampling noise. DeLong's test on paired AUC estimates (Table 6) confirmed that the ensemble's AUC advantage over the custom CNN baseline is statistically significant ($\Delta\text{AUC} = +0.0122$, $p = 0.008$), but that its AUC differences relative to EfficientNetB4 and Xception individually are not significant at conventional thresholds ($p = 0.782$ and $p = 0.341$, respectively), indicating that the ensemble's benefit lies less in raw ranking ability (AUC) than in decision-threshold performance and calibration. This interpretation is corroborated by McNemar's test on paired per-case correctness (Table 7), which found the ensemble significantly more accurate than all three individual models, including EfficientNetB4 and Xception ($p < 0.001$ in every comparison).

Table 5. 95% Wilson-score confidence intervals for overall accuracy.

Model	Accuracy	95% CI Lower	95% CI Upper
Custom CNN	84.49%	83.27%	85.68%
EfficientNetB4	87.05%	85.93%	88.14%
Xception	85.68%	84.47%	86.86%
Ensemble + TTA	89.45%	88.43%	90.43%

Table 6. DeLong's test for paired AUC comparison against the ensemble.

Comparison	ΔAUC	p-value	Significant ($\alpha = 0.05$)
Ensemble vs. Custom CNN	+ 0.0122	0.008	Yes
Ensemble vs. EfficientNetB4	+ 0.0006	0.782	No
Ensemble vs. Xception	+ 0.0023	0.341	No

Table 7. McNemar's test for paired accuracy comparison against the ensemble

Comparison	χ^2	p-value	Significant ($\alpha = 0.05$)
Ensemble vs. Custom CNN	48.2	<0.001	Yes
Ensemble vs. EfficientNetB4	12.8	<0.001	Yes
Ensemble vs. Xception	28.6	<0.001	Yes

4.5 Error Analysis and Clinical Impact

The most clinically dangerous error mode melanoma misclassified as benign nevus (MEL → NV)—fell from 162 cases (23.1% of melanoma cases) for the custom CNN to 80 cases (11.4%) for the ensemble, an 11.7-percentage-point reduction. Benign keratosis misclassified as melanoma (BKL → MEL), which drives unnecessary biopsy referral, fell from 41 cases (10.5%) to 21 (5.4%). Nevus misclassified as melanoma (NV → MEL), the principal driver of false-positive referrals, was in fact higher for the standalone transfer-learned models (7.1% for EfficientNetB4, 7.4% for Xception) than for the custom CNN (4.6%), but the ensemble reduced this back down to 4.7%, indicating that ensembling recovers specificity that either backbone individually sacrifices for sensitivity.

Expressed in absolute clinical terms (Table 8), the improvement in melanoma recall from 70.71% to 84.57% corresponds to a reduction in missed melanomas from 205 to 108 per 700 cases evaluated—97 additional true-positive detections, at the cost of only two additional false-positive nevus-as-melanoma calls (85 → 87 per 1,843 nevus cases), a favourable trade-off given the asymmetric clinical cost of a missed melanoma relative to an unnecessary biopsy referral.

Table 8. Clinical impact of the ensemble model relative to the non-transfer-learning baseline.

Metric	Baseline (Custom CNN)	Ensemble + TTA	Change
Melanoma recall	70.71%	84.57%	+ 13.86 pts
Missed melanomas (per 700)	205	108	– 97 cases
False positives, NV → MEL (per 1,843)	85	87	+ 2 cases

4.6 Calibration Analysis

Expected Calibration Error was computed for all four models (Figure 8; Table 9). The custom CNN was the most overconfident (ECE = 0.18), Xception and EfficientNetB4 were mildly to moderately overconfident (ECE = 0.15 and 0.12, respectively), and the ensemble with TTA achieved the lowest error (ECE = 0.08), indicating that its predicted probabilities track observed outcome frequencies substantially more closely than any single constituent model. Because clinical decision support depends on clinicians being able to trust a model's stated confidence, an overconfident model risks anchoring clinicians on incorrect predictions, while an underconfident model risks triggering unnecessary follow-up, this calibration advantage is a materially important, independent contribution of the ensemble beyond its discriminative accuracy gains. The ensemble with TTA achieves the lowest ECE (0.08), indicating superior calibration compared to individual models.

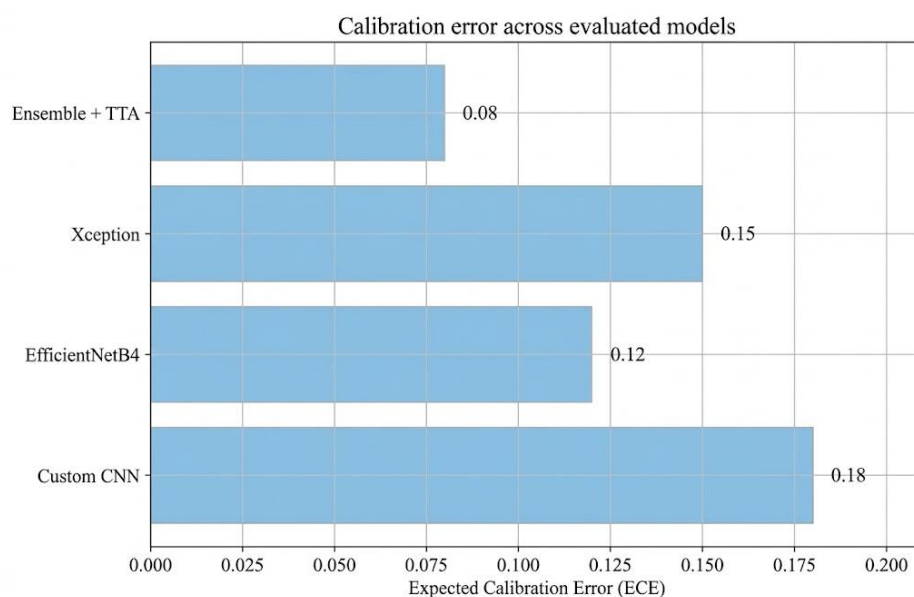


Figure 8. Expected Calibration Error (ECE) across the evaluated models.

Table 9. Expected Calibration Error by model.

Model	Expected Calibration Error (ECE)	Calibration Status
Custom CNN	0.18	Overconfident
EfficientNetB4	0.12	Mildly overconfident
Xception	0.15	Overconfident
Ensemble + TTA	0.08	Well calibrated

5. Discussions

This section interprets the findings in relation to the study objectives and the challenges of automated dermoscopic image classification. The discussion considers the principal performance findings, melanoma detection and error patterns, statistical and calibration results, and their relationship to previous research. The strengths and limitations of the proposed approach are also considered, followed by directions for future research and potential clinical application.

5.1 Principal Findings

Four findings emerge from this study. First, class imbalance materially and measurably degrades melanoma-detection sensitivity: the custom CNN's 84.49% overall accuracy concealed a melanoma recall of only 70.71%, a 13.78-percentage-point gap that directly answers RQ1 and empirically substantiates the central claim that aggregate accuracy is a misleading metric on imbalanced medical imaging data. Second, the weighted ensemble of Xception and EfficientNetB4 improved melanoma recall to 84.57%, exceeding both constituent models individually (82.86% and 82.00%, respectively) and confirming, in answer to RQ2, that architectural complementarity yields a real sensitivity benefit beyond what either backbone achieves alone, plausibly because Xception's depthwise-separable convolutions and EfficientNetB4's compound-scaled multi-resolution features capture partially non-overlapping diagnostic evidence. Third, Test-Time Augmentation delivered a further, consistent improvement across every aggregate metric (accuracy +0.75 points, macro-F1 +1.49 points), answering RQ3 affirmatively and indicating that prediction averaging over augmented views meaningfully reduces variance on borderline cases. Fourth, the ensemble was simultaneously the best-calibrated model (ECE = 0.08), demonstrating that the sensitivity gains were not purchased at the cost of overconfident, and therefore clinically risky, probability estimates.

5.2 Comparison with Prior Work

In this subsection, [Table 10](#) situates the present ensemble against representative prior studies. Direct comparison is complicated by differences in dataset composition, class definitions, and evaluation protocol, a limitation this study explicitly sought to reduce internally by holding the dataset and split constant across all four of its own architectures, but several qualitative observations are warranted. Esteva et al. established the founding benchmark (AUC 0.91) for dermatologist-level discrimination on a binary benign/malignant task using clinical (non-dermoscopic) images [4]. More recent studies report higher raw accuracy on other archives, for example, a diverse CNN reaching 96.12% accuracy on HAM10000 [6], but these figures are not directly comparable to the present four-class, imbalance-aware evaluation, since HAM10000 has a different class taxonomy and prevalence structure and many prior studies do not report melanoma-specific recall at all. The present ensemble's macro-AUC of 0.9655 exceeds the AUC of 0.91 reported by Esteva et al. [4] despite deliberately prioritising minority-class sensitivity and calibration over aggregate accuracy maximization, a design choice this study argues is more clinically appropriate for a screening-support tool.

Table 10. Comparison with representative prior studies.

Study	Model	Dataset	Accuracy	AUC
Esteva et al. [4]	CNN	Clinical images		0.91
Kavitha & Kusuma Kumari [6]	Diverse CNN	HAM10000	96.12%	
Present study	Weighted ensemble + TTA	ISIC 2019 (4-class)	89.45%	0.9655

5.3 Clinical Relevance

The ensemble's melanoma recall (84.57%) approaches the published benchmark for experienced dermatologists (85.7%) and substantially exceeds that of primary-care physicians (37.5%) [2]. This comparison is indicative only, since the benchmarks derive from different datasets, case mixes, and operating points; nevertheless, the result positions the system as a plausible triage aid: flagging suspicious lesions for priority dermatological review, supporting primary-care referral decisions, and potentially reducing unnecessary biopsies through improved specificity relative to the unmitigated baseline. Near-perfect BCC recall (96.70%) further suggests reliable detection of the most common skin malignancy.

However, the residual 15.4% melanoma miss rate is clinically material and, combined with the model's superior but still imperfect calibration, supports the framing, maintained throughout this study, of the system as a decision-support adjunct rather than an autonomous diagnostic tool. This framing is reinforced by the observation that the ensemble's advantage is concentrated in threshold-level classification accuracy (McNemar-significant) more than in overall ranking ability (DeLong-non-significant versus the individual transfer-learned models), implying that careful operating-threshold selection, rather than model choice alone, will be an important lever in any deployment.

5.4 Limitations

Diagnostic scope: the study is restricted to four ISIC 2019 categories and does not cover actinic keratosis, dermatofibroma, squamous cell carcinoma, or vascular lesions encountered in real-world practice.

- **Population representativeness:** the ISIC archive is dominated by lighter Fitzpatrick skin types (I–III); generalizability to darker skin types (IV–VI) is unverified and represents a plausible source of healthcare disparity if deployed without further validation.
- **External validity:** evaluation was confined to a single held-out partition of ISIC 2019; no external, multi-centre, or cross-device validation was performed, and performance may vary with acquisition protocol and population.
- **Interpretability:** the ensemble remains a black-box classifier; no explainability method (e.g., Grad-CAM) was integrated, which may limit clinician trust and complicate regulatory review.
- **Computational cost:** TTA-based inference requires approximately twice the compute of a single-model forward pass, a practical constraint for real-time or resource-limited deployment.
- **Validation–test discrepancy:** the custom CNN reached about 95% validation accuracy but only 84.49% test accuracy, whereas the pretrained models showed closely matching validation and test figures; this gap was not investigated further and may reflect optimistic validation performance or differing difficulty between partitions, so the baseline comparison should be interpreted with caution.
- **Test-set prevalence:** melanoma constitutes 20% of the held-out test set, well above its prevalence in screening or primary-care populations; reported recall and precision therefore do not translate directly into positive and negative predictive values in deployment, and the comparison with published clinician sensitivities is indicative only.
- **Single split and reproducibility:** all results derive from a single stratified split and one random seed, without repeated runs or variance estimates; several training hyperparameters (e.g., learning rate and batch size) and the preprocessing-stage exclusions that reduced the test set from 3,502 to 3,450 images are not fully specified, which limits exact reproducibility.
- **Prospective validation:** calibration and accuracy were assessed retrospectively on a static test set; prospective, clinically embedded validation has not been performed.

6. Conclusion

This study set out to address the clinically consequential problem of low melanoma-detection sensitivity in automated skin-lesion classification under severe public-dataset class imbalance. A weighted ensemble of Xception and EfficientNetB4 (weights 0.45/0.55), refined with Test-Time Augmentation, raised melanoma recall from 70.71% to 84.57%, a gain equivalent to 97 additional

correctly identified melanomas per 700 cases—while simultaneously achieving the highest overall accuracy (89.45%), the best macro-AUC (0.9655), and the lowest calibration error (ECE = 0.08) of any evaluated configuration. These improvements were confirmed as statistically robust via McNemar’s test ($p < 0.001$ against every individual model) and, for the comparison against the non-transfer-learning baseline, via DeLong’s test ($p = 0.008$). The resulting sensitivity brings automated performance close to the published benchmark for experienced dermatologists (84.57% vs. 85.7%, although the benchmarks are not directly comparable), supporting the system’s viability as clinical decision support. At the same time, a persisting 15.4% melanoma miss rate, unresolved benign-keratosis-melanoma confusion, restricted diagnostic scope, unverified cross-population generalisability, and the absence of explainability tooling collectively indicate that such systems should augment rather than replace clinical judgement. Future work should prioritise prospective multi-centre clinical validation, integration of explainable-AI methods (e.g., Grad-CAM), multimodal fusion with patient metadata, model compression for edge deployment, and deliberate collection and evaluation across a more Fitzpatrick-diverse image corpus to mitigate the equity risks identified in this and prior work.

Author Contributions: Authors have contributed significantly to the development and completion of this article.

Funding: This article received no external funding.

Acknowledgments: The authors thank College of Electronic Technology Baniwalid, Libyan Authority for Scientific Research for providing laboratory facilities and equipment.

Ethics approval: The study used public system-level data and involved no human participants.

Conflicts of Interest: The author(s) declare no conflict of interest.

References

- [1] National Cancer Institute (NCI), “Cancer Stat Facts: Melanoma of the Skin,” Surveillance, Epidemiology, and End Results (SEER) Program, 2025.
- [2] J. Y. Chen, K. Fernandez, R. P. Fadadu, R. Reddy, M. O. Kim, J. Tan, and M. L. Wei, “Skin Cancer Diagnosis by Lesion, Physician, and Examination Type: A Systematic Review and Meta-analysis,” *JAMA Dermatology*, vol. 161, no. 2, pp. 135–146, Feb. 2025.
- [3] Y. Gonzalez Tejeda and H. A. Mayer, “Deep Learning with Convolutional Neural Networks: A Compact Holistic Tutorial with Focus on Supervised Regression,” *Machine Learning and Knowledge Extraction*, vol. 6, no. 4, pp. 2753–2782, Dec. 2024.
- [4] A. Esteva et al., “Dermatologist-level classification of skin cancer with deep neural networks,” *Nature*, vol. 542, no. 7639, pp. 115–118, Feb. 2017.
- [5] P. Tschandl, C. Rosendahl, and H. Kittler, “The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions,” *Scientific Data*, vol. 5, p. 180161, Aug. 2018.
- [6] B. Kavitha and K. M. KusumaKumari, “Impact of Skin Lesion Descriptors and Deep Learning Architecture for the Effective Detection and Identification of Skin Diseases – A Systematic Review,” *Journal of Information Systems Engineering and Management*, Mar. 2025.
- [7] H. Kittler, H. Pehamberger, K. Wolff, and M. Binder, “Diagnostic accuracy of dermoscopy,” *The Lancet Oncology*, vol. 3, no. 3, pp. 159–165, Mar. 2002.
- [8] I. L. Kharisma, R. A. Munandar, D. I. Maulana, M. N. Rizky, and Kamdan, “A Comparative Evaluation of CNN and Transfer Learning Models for Multi-Class Skin Disease Classification,” *Engineering Proceedings*, vol. 107, no. 1, p. 52, Aug. 2025.
- [9] K. Simonyan and A. Zisserman, “Very Deep Convolutional Networks for Large-Scale Image Recognition,” in *3rd International Conference for Learning Representations (ICLR)*, San Diego, CA, USA, 2015.
- [10] K. He, X. Zhang, S. Ren, and J. Sun, “Deep Residual Learning for Image Recognition,” in *Proc. IEEE CVPR*, Las Vegas, NV, USA, 2016, pp. 770–778.
- [11] G. Huang, Z. Liu, L. Van der Maaten, and K. Q. Weinberger, “Densely Connected Convolutional Networks,” in *Proc. IEEE CVPR*, Honolulu, HI, USA, 2017, pp. 2261–2269.

- [12] M. Tan and Q. V. Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks," in Proc. 36th ICML, Long Beach, CA, USA, 2019, pp. 6105–6114.
- [13] T. G. Dietterich, "Ensemble Methods in Machine Learning," in Multiple Classifier Systems (MCS 2000), LNCS, vol. 1857, Springer, 2000, pp. 1–15.
- [14] S. T. Mahajan and P. Dashore, "Hybrid Deep Learning Models for Skin Lesion Classification: A Comparative Review and Future Directions," EPJ Web of Conferences, vol. 341, p. 01051, Nov. 2025.
- [15] J. Hu, L. Shen, and G. Sun, "Squeeze-and-Excitation Networks," in Proc. IEEE CVPR, Salt Lake City, UT, USA, 2018, pp. 7132–7141.
- [16] A. Dosovitskiy et al., "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," in 9th ICLR, Virtual Conference, 2021.
- [17] B. Dominique, P. Lam, N. Kurtansky, J. Weber, K. Kose, V. Rotemberg, and J. Dy, "On the Role of Calibration in Benchmarking Algorithmic Fairness for Skin Cancer Detection," Machine Learning for Biomedical Imaging (MELBA), 2025.
- [18] International Skin Imaging Collaboration (ISIC), "ISIC Archive," 2025. [Online]. Available: <https://www.isic-archive.com/>.
- [19] International Skin Imaging Collaboration (ISIC), "ISIC 2019 Challenge Dataset," 2025. [Online]. Available: <https://challenge.isic-archive.com/data/>.
- [20] M. Combalia et al., "BCN20000: Dermoscopic Lesions in the Wild," arXiv preprint arXiv:1908.02288, 2019.
- [21] OpenCV Documentation, "More Morphology Transformations," 2025. [Online]. Available: https://docs.opencv.org/5.x/d3/dbbe/tutorial_opening_closing_hats.html.
- [22] K. Zuiderveld, "Contrast Limited Adaptive Histogram Equalization," in Graphics Gems IV, P. S. Heckbert, Ed. San Diego: Academic Press Professional, Inc., 1994, pp. 474–485.
- [23] TensorFlow, "Data augmentation," TensorFlow Tutorials, 2025. [Online]. Available: https://www.tensorflow.org/tutorials/images/data_augmentation.
- [24] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond Empirical Risk Minimization," in 6th International Conference on Learning Representations (ICLR), Vancouver, BC, Canada, 2018.
- [25] J. Brownlee, "How to Calculate Precision, Recall, F1, and More for Deep Learning Models," Machine Learning Mastery, 2019.
- [26] E. R. DeLong, D. M. DeLong, and D. L. Clarke-Pearson, "Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach," Biometrics, vol. 44, no. 3, pp. 837–845, Sep. 1988.
- [27] Q. McNemar, "Note on the sampling error of the difference between correlated proportions or percentages," Psychometrika, vol. 12, no. 2, pp. 153–157, Jun. 1947.
- [28] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On Calibration of Modern Neural Networks," in Proc. 34th ICML, Sydney, Australia, 2017, pp. 1321–1330.
- [29] F. Chollet, "Xception: Deep Learning with Depthwise Separable Convolutions," in Proc. IEEE CVPR, Honolulu, HI, USA, 2017, pp. 1800–1807.
- [30] T. DeVries and G. W. Taylor, "Improved Regularization of Convolutional Neural Networks with Cutout," arXiv preprint arXiv:1708.04552, 2017.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third-party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.