GMR
The Original

DERMAEDGE-X: AN EFFICIENCY-ORIENTED AND CLASS-IMBALANCE-AWARE DEEP LEARNING FRAMEWORK FOR MULTICLASS SKIN LESION CLASSIFICATION

Rajneesh Kumar[1], V Senthilkumar[2], Gripsy Paul Mannickathan[3], Saju A[4], Gourav Kalra[5], Sudeep Kumar T[6], Asesh Kumar Tripathy[7], Divya Saleela[8]

[1]Assistant Professor, Department of Computer Science and Engineering,
Faculty of Engineering and Technology, SRM Institute of Science and Technology, DELHI-NCR Campus, Delhi-Meerut Road, Modinagar, Uttar Pradesh, India.
[2]Professor, SRM TRP Engineering College, Tamil Nadu, India
[3]Assistant Professor, CMR University, Karnataka, India
[4]Professor, Albertian Institute of Science and Technology, Kerala, India
[5]Associate Professor, School of Engineering and Technology, CGC University, Punjab, India
[6]Assistant Professor, Siddaganga Institute of Technology, Karnataka, India
[7]Associate Professor, Koneru Lakshmaiah Education Foundation, Andhra Pradesh, India
[8]Assistant Professor, College of Engineering Aranmula, Kerala, India

ABSTRACT
Automated dermoscopic image classification remains challenging because of severe class imbalance, inter-class visual similarity and the computational demands of contemporary deep learning models. This study presents DermaEdge-X, an efficiency-oriented framework for seven-class skin lesion classification using lesion-grouped data partitioning, transfer learning, class-weighted optimisation, label smoothing, image augmentation, validation-guided checkpoint selection and test-time augmentation. The framework was evaluated on a lesion-disjoint held-out test set using predictive, calibration, robustness and computational metrics. DermaEdge-X achieved 88.19% accuracy, 73.99% balanced accuracy, a macro F1-score of 76.98%, a macro receiver operating characteristic area under the curve of 0.9695 and a macro precision–recall area under the curve of 0.8610. The model contained 51.19 million parameters, required 51.04 GFLOPs per image and achieved a median inference latency of 23.85 ms per image on an NVIDIA Tesla T4 GPU. The proposed framework provides a reproducible basis for developing class-imbalance-aware and deployment-conscious skin lesion classification systems for computer-assisted dermatological screening.

KEYWORDS: Dermoscopic image classification; efficient deep learning; transfer learning; melanoma detection.

1. INTRODUCTION

Skin cancer remains a major and increasing global health burden, with more than 1.5 million cases diagnosed worldwide and melanoma accounting for approximately 330,000 cases and almost 60,000 deaths in the latest comprehensive global assessment [1] [2]. Dermoscopy facilitates the non-invasive visualisation of subsurface skin structures that are not readily apparent during routine clinical examination and has therefore become an important tool for evaluating suspicious pigmented and non-pigmented lesions. However, dermoscopic diagnosis remains challenging because benign and malignant lesions often exhibit overlapping colours, borders, textures and morphological patterns. Melanoma, for example, may closely resemble a benign melanocytic naevus, while basal cell carcinoma, actinic keratosis, benign keratosis-like lesions, dermatofibroma and vascular lesions may also share similar visual characteristics. Diagnostic accuracy may be further influenced by clinician expertise, anatomical site, image quality, acquisition conditions and variations in skin pigmentation. Automated analysis of dermoscopic images therefore offers considerable potential to complement specialist assessment, improve diagnostic consistency and support the prioritisation of high-risk lesions for further clinical investigation.

Deep learning has substantially advanced automated skin lesion analysis by enabling discriminative visual representations to be learned directly from images. Convolutional neural networks have demonstrated strong capability in capturing local texture, colour and structural information, while transfer learning has made it possible to adapt models pretrained on large natural-image datasets to comparatively small medical imaging datasets [3]. More recent architectures have introduced improved feature hierarchies, multiscale processing and attention mechanisms. ConvNeXt-based models modernise conventional convolutional networks using architectural principles associated with vision transformers, while Swin Transformer V2 employs hierarchical shifted-window attention and improved positional modelling. EfficientNetV2, MaxViT and MobileNetV4 provide alternative balances between representational capacity and computational efficiency. These developments demonstrate the

potential of modern visual architectures for dermoscopic image classification while also highlighting the need to assess their computational requirements alongside predictive performance.

Despite this progress, important methodological and practical limitations remain. Dermoscopic datasets are typically highly imbalanced, with melanocytic naevi substantially outnumbering clinically important minority classes such as dermatofibroma and vascular lesions [4]. Models trained without appropriate imbalance mitigation may consequently favour majority classes and achieve deceptively high overall accuracy while performing poorly on less frequent categories [5]. Furthermore, random image-level partitioning can place multiple images of the same lesion across training and evaluation sets, introducing information leakage and producing overly optimistic estimates of generalisation. Many studies also report only accuracy or weighted performance measures without adequately considering balanced accuracy, macro-averaged metrics, agreement measures or precision–recall performance. In addition, increasingly complex architectures may require substantial memory, high floating-point operation counts and prolonged inference times, restricting their suitability for mobile, edge-based and resource-constrained clinical settings. There is therefore a need for an evaluation framework that jointly addresses lesion-level data separation, class imbalance, predictive reliability and computational requirements.

To address these limitations, this study proposes DermaEdge-X, an efficiency-oriented and class-imbalance-aware framework for seven-class dermoscopic skin lesion classification. The principal contributions of this work are: (i) a lesion-grouped partitioning strategy designed to minimise information leakage across the training, validation and held-out test sets; (ii) a bounded square-root inverse-frequency weighting scheme that reduces majority-class dominance without excessively amplifying rare-class contributions; (iii) a reproducible learning pipeline incorporating transfer learning, image augmentation, mixup, label smoothing, validation-guided checkpoint selection and test-time augmentation; (iv) comprehensive evaluation using accuracy, balanced accuracy, macro and weighted precision, recall and F1-score, Matthews correlation coefficient, Cohen's kappa, receiver operating characteristic area under the curve and precision–recall area under the curve; and (v) analysis of probability calibration, robustness under controlled image perturbations and computational characteristics. The study additionally evaluates parameter count, floating-point operations, inference latency, model size and training time to characterise the relationship between predictive performance and deployment-related computational constraints.

The remainder of this paper is organised as follows. Section 2 reviews previous research on automated dermoscopic image classification, class-imbalance mitigation and computationally efficient deep learning architectures. Section 3 describes the dataset, preprocessing procedures, lesion-grouped partitioning strategy, DermaEdge-X architecture, optimisation procedure and evaluation measures. Section 4 presents the training behaviour, held-out test performance, class-wise results, calibration, robustness, computational characteristics and interpretability analysis. Section 5 discusses the principal findings, clinical relevance and study limitations. Finally, Section 6 concludes the paper and identifies directions for external validation and future research.

2. LITERATURE REVIEW

Automated skin lesion classification has evolved from conventional machine-learning pipelines based on manually engineered colour, texture and shape descriptors to end-to-end deep learning systems capable of learning discriminative representations directly from dermoscopic images. Earlier approaches typically combined image preprocessing, lesion segmentation, handcrafted feature extraction and classifiers such as support vector machines, random forests and k-nearest neighbours [6] [7]. Although these methods provided interpretable feature pipelines, their performance depended heavily on the quality of lesion segmentation and the suitability of predefined descriptors. Their limited capacity to represent complex morphological variation also reduced their robustness across visually similar diagnostic categories.

The introduction of convolutional neural networks substantially improved automated dermoscopic image analysis. Architectures such as VGG, ResNet, Inception, DenseNet, MobileNet and EfficientNet enabled hierarchical representations of pigmentation, borders, texture and lesion structure to be learned directly from images [8] [9]. Transfer learning became particularly important because medical imaging datasets are generally smaller than large natural-image repositories. By adapting ImageNet-pretrained models, researchers were able to reduce training requirements and improve convergence [10]. Subsequent studies incorporated extensive augmentation, attention mechanisms, ensemble learning, lesion segmentation and clinical metadata fusion to improve classification performance. However, many of these systems relied on computationally demanding backbones or multi-model ensembles, limiting their suitability for efficient clinical deployment.

Recent research has increasingly focused on modern convolutional and transformer-based architectures [11] [12]. ConvNeXt modernised conventional convolutional networks using design principles associated with vision transformers, while ConvNeXt-V2 further introduced masked autoencoding and global response normalisation to strengthen representation learning. Vision transformers model long-range spatial relationships through self-attention, allowing information from distant image regions to be integrated. Swin Transformer V2 improves computational tractability through hierarchical shifted-window attention, whereas MaxViT combines local window attention with global grid attention. EfficientNetV2 balances representational capacity and training efficiency through architectural scaling and progressive learning, while MobileNetV4 targets low-latency deployment through hardware-aware design. These architectures illustrate the range of predictive and computational trade-offs available for skin lesion classification and motivate evaluation protocols that consider both performance and resource requirements.

Class imbalance remains a major challenge in multiclass dermoscopic image classification. Common lesion categories frequently dominate available training data, while diagnostically important classes may contain relatively few examples. Models optimised without suitable imbalance mitigation can therefore favour majority categories and produce high overall accuracy despite inadequate minority-class recognition. Existing studies have addressed this problem through oversampling, targeted augmentation, synthetic-image generation, focal loss, class-weighted cross-entropy, balanced sampling and cost-sensitive learning. Although these methods can improve rare-class sensitivity, excessive oversampling may promote memorisation, while unrestricted inverse-frequency weighting may assign excessively large penalties to rare classes and destabilise optimisation. More controlled weighting strategies are therefore required to increase minority-class influence without introducing disproportionate gradient contributions.

The reliability of reported classification performance also depends strongly on the data-partitioning strategy. Dermoscopic repositories may contain multiple images corresponding to the same lesion or patient. Random image-level splitting can consequently place highly related images across the training, validation and test sets, creating information leakage and inflating estimates of generalisation. Nevertheless, many studies provide insufficient information about whether lesion or patient identities were kept disjoint during partitioning. Some investigations also rely on a single development split without a clearly separated held-out test set. Lesion-grouped partitioning is therefore important for ensuring that evaluation reflects performance on genuinely unseen lesions rather than recognition of lesion-specific characteristics encountered during training.

Evaluation practices in the literature are similarly inconsistent. Overall accuracy remains the most frequently reported measure, despite its limited informativeness under severe class imbalance. Weighted precision, recall and F1-score are also influenced by majority classes and may conceal weak performance on under-represented categories. Balanced accuracy and macro-averaged precision, recall and F1-score provide a more equitable assessment by assigning equal importance to each diagnostic class. Matthews correlation coefficient and Cohen's kappa offer complementary measures of overall classification quality and agreement beyond chance. Receiver operating characteristic area under the curve assesses ranking discrimination, while precision–recall area under the curve is particularly relevant when class distributions are skewed. A comprehensive evaluation should therefore combine overall, class-balanced, agreement and discrimination measures rather than relying on a single summary statistic.

Another limitation of existing research is the limited attention given to computational efficiency. Skin lesion classifiers are commonly ranked according to accuracy or F1-score, while parameter count, floating-point operations, model size, inference latency and training time are either omitted or reported under different hardware and preprocessing conditions. This makes it difficult to determine whether small predictive improvements justify substantial increases in memory use or computational cost. Such considerations are especially important for mobile dermoscopy, edge-based screening and resource-constrained clinical environments. Classification frameworks should therefore be evaluated according to both predictive effectiveness and deployment-related computational requirements.

The literature reveals four principal gaps. First, many studies do not explicitly adopt lesion-grouped partitioning, creating a risk of information leakage. Second, class imbalance is often addressed through basic resampling or unrestricted weighting without controlling the influence of rare categories. Third, evaluation commonly prioritises accuracy and weighted metrics, providing limited evidence of balanced multiclass performance. Fourth, computational characteristics such as parameter count, floating-point operations, inference latency, model size and training time are frequently omitted, making it difficult to assess deployment-related constraints. The present study addresses these gaps through DermaEdge-X, which combines lesion-grouped training, validation and held-out test partitions with bounded square-root inverse-frequency weighting, comprehensive class-balanced evaluation, probability calibration, robustness analysis and computational profiling. The study evaluates predictive performance alongside parameter count, floating-point operations, inference latency, model size and training time to provide a deployment-conscious assessment of seven-class skin lesion classification.

3. MATERIALS AND METHODS

This section describes the proposed DermaEdge-X framework for seven-class dermoscopic image classification. The methodological pipeline comprised lesion-aware data separation, image preprocessing and augmentation, transfer-learning-based feature extraction, class-imbalance-aware optimisation, validation-guided model selection and test-time prediction. Figure 1 illustrates the complete architecture and information flow of the proposed framework.

3.1 Lesion-Aware Data Partitioning

The image collection was partitioned into training, validation and held-out lesion-disjoint test subsets using lesion identifiers as grouping variables. All images associated with the same lesion were retained within a single subset, thereby preventing lesion-specific visual information from being shared across model-development and evaluation partitions. The training subset was used for parameter optimisation, the validation subset was used for checkpoint selection and early stopping, and the held-out lesion-disjoint test subset remained isolated until the final evaluation.

3.2 Image Preprocessing and Augmentation

All dermoscopic images were converted to three-channel RGB format and resized to 384 × 384 pixels. The training images were processed using stochastic augmentation and mixup to increase visual diversity and reduce

overfitting. Mixup was applied during fine-tuning epochs 3–30 and disabled during the final five epochs. Normalisation was performed consistently across all subsets using the preprocessing configuration associated with the pretrained feature extractor.

Training images were loaded in randomly shuffled mini-batches, while validation and test images were processed sequentially. Augmentation was restricted to the training phase to ensure that model selection and final evaluation were conducted using reproducible image representations.

3.3 DermaEdge-X Architecture

DermaEdge-X was developed as a transfer-learning-based multiclass classification framework, as illustrated in Figure 1. The architecture consists of a pretrained visual feature extractor followed by a task-specific classification head. Given an input dermoscopic image, the feature extractor generates hierarchical representations that capture lesion colour, texture, border characteristics and internal morphological structure. The resulting feature representation is passed to the classification head, which produces seven output logits corresponding to the diagnostic categories. These logits are converted into posterior class probabilities using the softmax function.

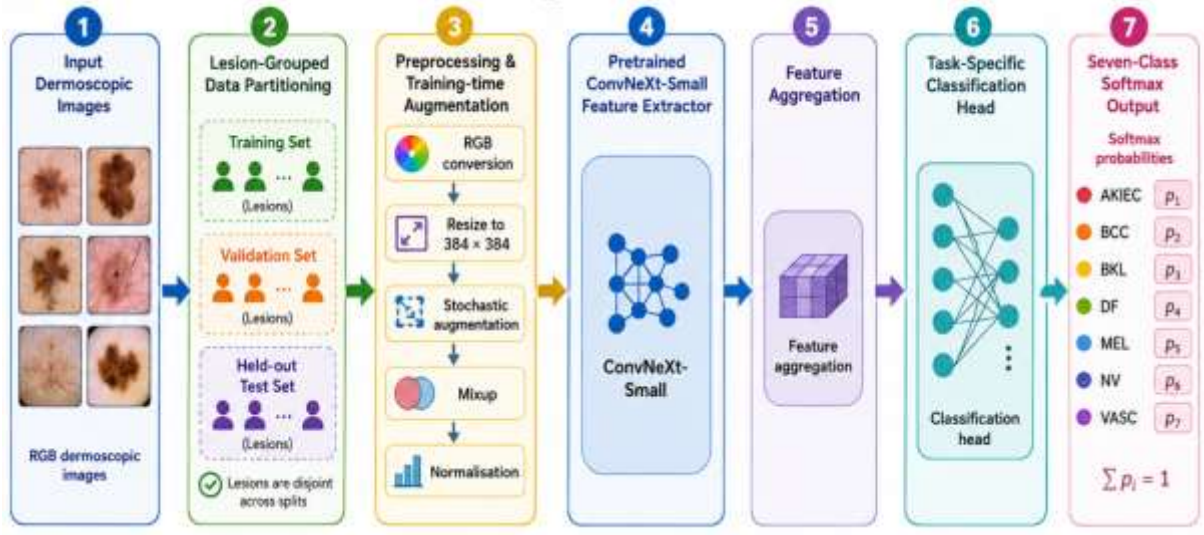


Figure 1. Overall workflow of the proposed DermaEdge-X framework, showing lesion-grouped data partitioning, image preprocessing and training-time augmentation, pretrained ConvNeXt-Small feature extraction, feature aggregation, task-specific seven-class classification and softmax probability output.

3.4 Class-Imbalance-Aware Optimisation

Class imbalance was addressed using bounded square-root inverse-frequency weighting. For class c, the preliminary weight was calculated as

$$w_c = \sqrt{\frac{n_{max}}{\max(n_c,1)}} \qquad (1)$$

where $n_c$denotes the number of training samples in class c, and $n_{\max}$denotes the number of samples in the largest class.

The preliminary weights were normalised by their mean:

$$w_c' = \frac{w_c}{\frac{1}{C}\sum_{j=1}^{C} w_j} \qquad (2)$$

where $C$is the total number of classes. To prevent excessively large or small gradient contributions, the normalised weights were restricted to the interval $[0.65, 2.00]$:

$$\tilde{w}_c = \text{clip}(w_c', 0.65, 2.00) \quad (3)$$

The resulting weights were incorporated into categorical cross-entropy with label smoothing:

$$\mathcal{L} = -\sum_{c=1}^{C} \tilde{w}_c \; y_c^{(\varepsilon)} \log(p_c) \qquad (4)$$

where $p_c$is the predicted probability for class c, and $y_c^{(\varepsilon)}$is the label-smoothed target distribution. The label-smoothing coefficient was set to 0.04. This formulation increased the contribution of under-represented categories while limiting unstable optimisation caused by extreme inverse-frequency weights.

3.5 Model Training and Selection

The model was optimised using AdamW with an initial learning rate of $3 \times 10^{-5}$and a weight-decay coefficient of 8×10^{-5}. Training comprised a five-epoch warm-up phase followed by 35 epochs of end-to-end fine-tuning. A

cosine-annealing learning-rate schedule progressively reduced the learning rate to 3×10^{-7}. The checkpoint with the highest validation composite score was retained for final evaluation.
Mixed-precision computation was used during GPU training to reduce memory consumption and improve execution efficiency. Gradient norms were clipped at 3.0 to limit unstable parameter updates.
At the end of each epoch, the model was evaluated on the validation subset. Checkpoint selection was based on a composite criterion combining validation accuracy, balanced accuracy and macro F1-score:

$$S = 0.40A + 0.30B + 0.30F_{\text{macro}} \quad (5)$$

where A represents validation accuracy, B represents balanced accuracy, and F_{macro} represents macro F1-score. The parameter state associated with the highest composite score was retained. Early stopping was applied when the selection score failed to improve for six consecutive epochs.

3.6 Test-Time Prediction

During validation and held-out lesion-disjoint testing, predictions were generated from both the original image and a horizontally flipped version. The corresponding softmax probability vectors were combined as

$$\mathbf{p}_{\text{final}} = 0.70\mathbf{p}_{\text{original}} + 0.30\mathbf{p}_{\text{flipped}} \quad (6)$$

where $\mathbf{p}_{\text{original}}$ and $\mathbf{p}_{\text{flipped}}$ denote the predicted probability vectors from the original and transformed images, respectively. The final diagnostic label was assigned to the class with the highest fused probability.

3.7 Performance Evaluation

The retained validation-selected checkpoint was applied to the independent-test subset. Performance was quantified using accuracy, balanced accuracy, macro precision, macro recall, macro F1-score, weighted precision, weighted recall, weighted F1-score, Matthews correlation coefficient, and Cohen's kappa.
Multiclass discrimination was assessed using macro one-versus-rest receiver operating characteristic area under the curve and macro precision–recall area under the curve. Class-specific precision, recall, and F1-score were additionally calculated to examine predictive behaviour across individual diagnostic categories. All reported test measures were derived from the independent-test predictions generated by the retained checkpoint.

4. RESULTS

This section presents the optimisation behaviour, independent-test performance, class-wise classification results, discrimination and calibration characteristics, robustness under controlled image perturbations, computational requirements, and qualitative interpretability of DermaEdge-X. All test-set results were obtained using the validation-selected checkpoint under lesion-grouped evaluation.

4.1 Training and validation behaviour

DermaEdge-X was trained on 7,991 images and validated on 1,025 images. During the five-epoch warm-up phase, validation accuracy increased from 67.41% to 75.71%, balanced accuracy from 25.54% to 44.06%, and macro F1-score from 25.26% to 47.58%. Subsequent end-to-end fine-tuning produced sustained improvement across the validation measures. By fine-tuning epoch 8, the model had reached 88.29% accuracy, 80.04% balanced accuracy and a macro F1-score of 80.07%. The highest validation accuracy was observed at fine-tuning epoch 16, reaching 89.37%, whereas the strongest class-balanced performance occurred at fine-tuning epoch 29, with a balanced accuracy of 83.67% and a macro F1-score of 82.53%.
Training loss decreased from 0.8555 at the first fine-tuning epoch to 0.3353 at fine-tuning epoch 29. After approximately 20 fine-tuning epochs, the validation curves entered a stable region, with only minor oscillations in accuracy and macro F1-score. The training-loss trajectory is shown in Figure 2a, while Figure 2b presents the corresponding validation accuracy, balanced accuracy and macro F1-score across the warm-up and fine-tuning stages.

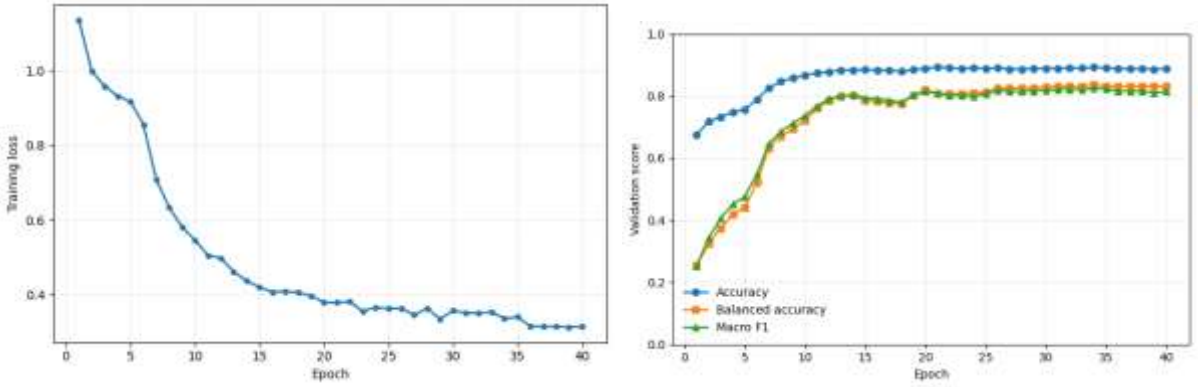


Figure 2. Training behaviour of DermaEdge-X: (a) training-loss trajectory and (b) validation accuracy, balanced accuracy and macro F1-score across the warm-up and fine-tuning stages.

4.2 Independent-test performance

The retained checkpoint was evaluated on an independent test set comprising 999 images from 747 unique lesions. DermaEdge-X achieved an accuracy of 88.19%, balanced accuracy of 73.99%, and macro F1-score of 76.98%. Weighted precision, weighted recall, and weighted F1-score were 88.48%, 88.19%, and 88.10%, respectively. Matthews correlation coefficient and Cohen's kappa were 0.7778 and 0.7776, indicating substantial agreement between predicted and reference labels.

The model also demonstrated strong multiclass discrimination, achieving a macro one-versus-rest ROC-AUC of 0.9695 and a macro PR-AUC of 0.8610. The complete independent-test performance of DermaEdge-X is illustrated in Figure 3.

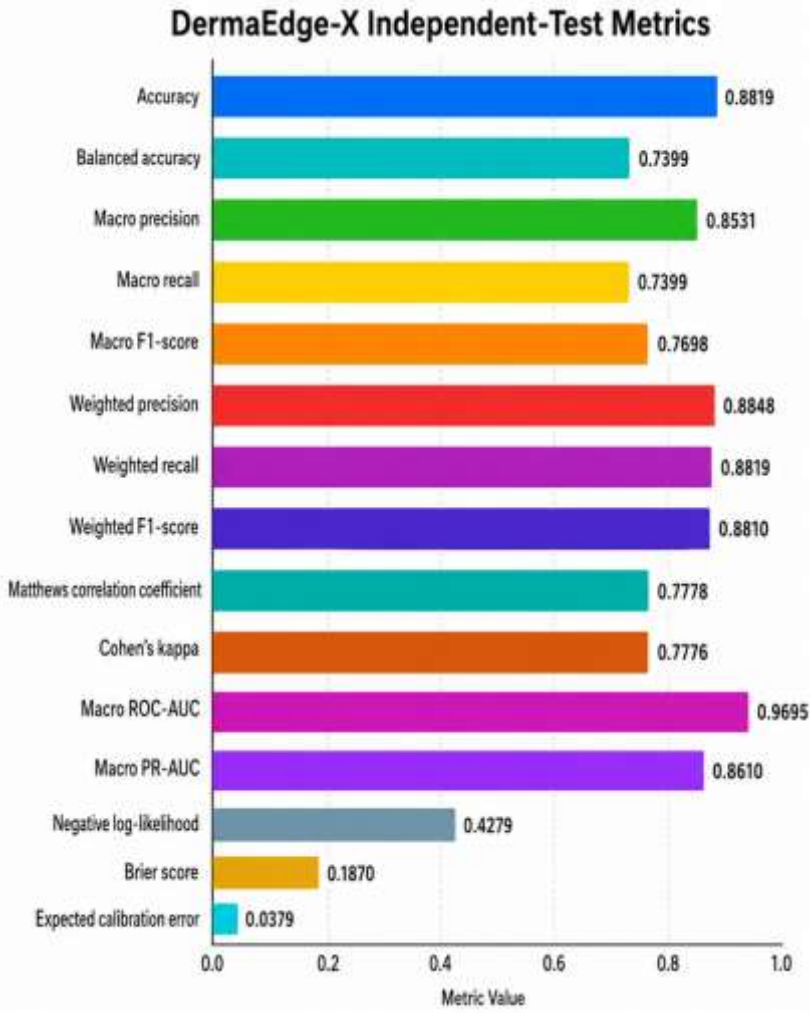


Figure 3. Independent-test performance metrics of DermaEdge-X.

4.3 Class-wise performance

Class-wise precision, recall, and F1-score are presented in Table 1. The strongest performance was obtained for melanocytic naevi, with precision, recall, and F1-score all equal to 0.9422. Vascular lesions also achieved high performance, with an F1-score of 0.9000. Basal cell carcinoma and benign keratosis-like lesions yielded F1-scores of 0.8506 and 0.8443, respectively.

Actinic keratosis and intraepithelial carcinoma achieved a precision of 0.8462 and recall of 0.6111, resulting in an F1-score of 0.7097. Melanoma remained more difficult to classify, with precision of 0.6182, recall of 0.6667, and F1-score of 0.6415. Dermatofibroma achieved precision of 1.0000 but recall of 0.3333, producing an F1-score of 0.5000; however, this estimate was based on only nine test samples.

Table 1. Class-wise independent-test performance

Class	Precision	Recall	F1-score	Support
AKIEC	0.8462	0.6111	0.7097	36
BCC	0.8409	0.8605	0.8506	43
BKL	0.8243	0.8652	0.8443	141
DF	1.0000	0.3333	0.5000	9
MEL	0.6182	0.6667	0.6415	102

Class	Precision	Recall	F1-score	Support
NV	0.9422	0.9422	0.9422	658
VASC	0.9000	0.9000	0.9000	10

The confusion matrix in Figure 4 showed that most predictions were concentrated along the principal diagonal. The most prominent residual errors involved confusion between melanoma and melanocytic naevi. High-confidence error analysis confirmed this pattern, with several naevi predicted as melanoma and several melanoma images predicted as naevi with confidence values above 0.95. Additional errors were observed between actinic keratosis and benign keratosis-like lesions.

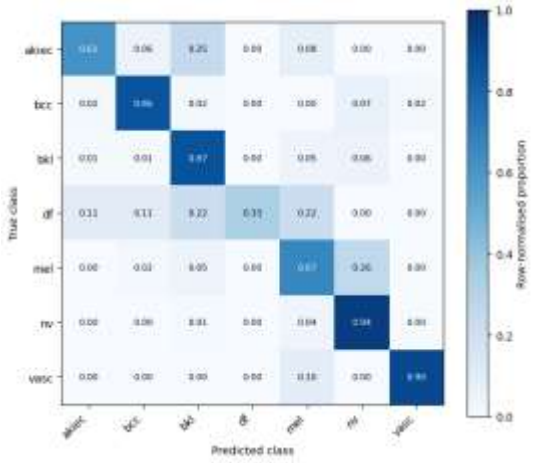
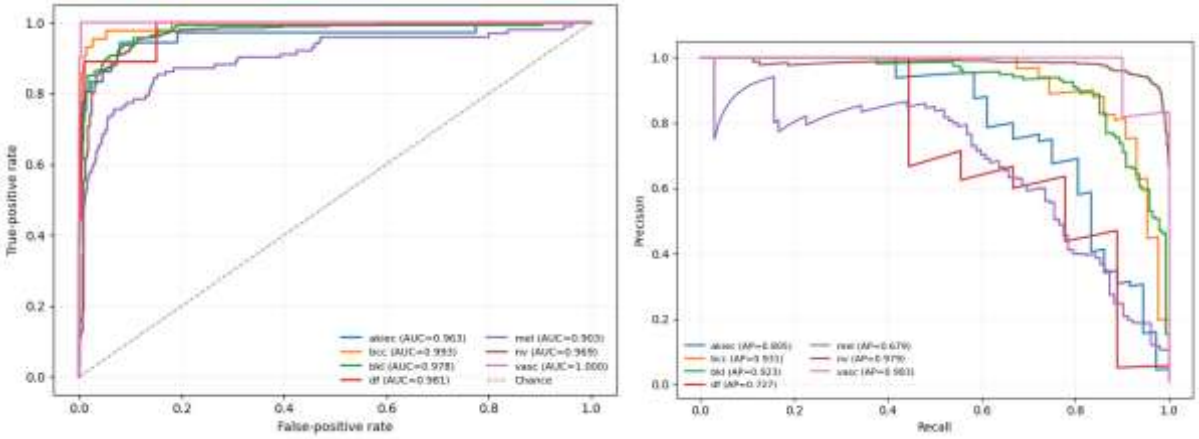


Figure 4. Row-normalised confusion matrix of DermaEdge-X on the independent test set.

4.4 Discrimination and calibration

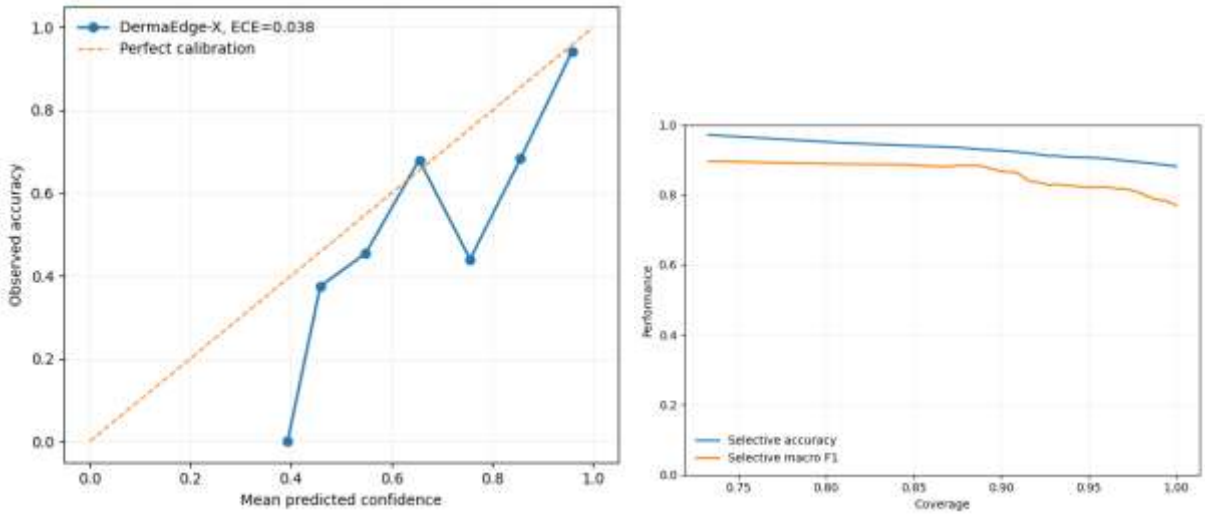
The multiclass ROC and precision–recall curves are shown in Figures 5a and 5b. The macro ROC-AUC of 0.9695 indicated strong separation among the seven diagnostic categories across classification thresholds. The macro PR-AUC of 0.8610 further demonstrated favourable discrimination under the imbalanced class distribution.



(a) (b)

Figure 5. Discrimination performance of DermaEdge-X on the independent test set: (a) one-versus-rest receiver operating characteristic curves and (b) class-wise precision–recall curves.

Temperature scaling identified an optimal temperature of 1.0658. Following calibration, the model achieved a negative log-likelihood of 0.4279, a Brier score of 0.1870 and an expected calibration error of 0.0379. The calibration curve is presented in Figure 6a, while Figure 6b shows the selective-prediction behaviour as low-confidence cases are progressively excluded.



(a) (b)

Figure 6. Reliability analysis of DermaEdge-X: (a) calibration curve following temperature scaling and (b) selective-prediction performance in terms of selective accuracy and selective macro F1-score as coverage decreases.

4.5 Robustness under image perturbations

Robustness was evaluated under blur, brightness variation, colour distortion and synthetic hair artefacts at five severity levels. The detailed robustness results are summarised in Table 2, and the corresponding trends are visualised in Figure 7. Across all perturbation types and severity levels, overall accuracy remained above 84%, although balanced accuracy and macro F1-score showed greater sensitivity to stronger image degradation.

Under blur, accuracy decreased from 87.19% at severity 1 to 84.48% at severity 5, while macro F1-score declined from 74.69% to 70.69%. Brightness perturbation produced accuracy values between 85.49% and 86.79%, with balanced accuracy decreasing from 69.41% to 65.79% as severity increased. Colour distortion had the largest adverse effect on class-balanced performance, with macro F1-score decreasing from 73.46% at severity 1 to 68.02% at severity 5. Under synthetic hair artefacts, accuracy ranged from 85.69% to 86.99%, while macro F1-score varied between 70.56% and 74.54%. Overall, the reduction in balanced accuracy and macro F1-score was generally greater than the reduction in accuracy, indicating that minority-class predictions were more sensitive to severe image perturbations.

Table 2. Robustness performance of DermaEdge-X across image perturbations

Perturbation	Accuracy range	Balanced accuracy range	Macro F1 range
Blur	0.8448–0.8719	0.6952–0.7275	0.7069–0.7469
Brightness	0.8549–0.8679	0.6579–0.6941	0.6951–0.7300
Colour	0.8569–0.8639	0.6339–0.7052	0.6696–0.7346
Hair	0.8569–0.8699	0.6676–0.7116	0.7056–0.7454

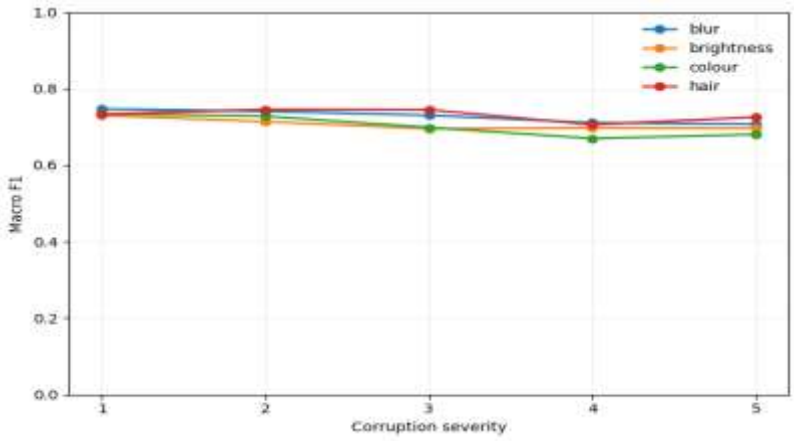


Figure 7. Effect of increasing perturbation severity on the macro F1-score of DermaEdge-X under blur, brightness, colour and synthetic hair artefacts.

4.6 Computational performance

As shown in Table 3, DermaEdge-X contained 51.19 million parameters and required 51.04 GFLOPs for inference on a single 384 × 384image. Median inference latency was 23.85 ms per image on an NVIDIA Tesla T4 GPU, corresponding to an approximate throughput of 41.9 images per second. The stored model size was 195.75 MB, and total training time was 12,777.79 s, equivalent to approximately 213 min.

Table 3. Computational characteristics of DermaEdge-X

Measure	Value
Parameters	51.19 M
Computational cost	51.04 GFLOPs/image
Median inference latency	23.85 ms/image
Approximate throughput	41.9 images/s
Model size	195.75 MB
Training time	12,777.79 s

4.7 Interpretability and error analysis

Gradient-weighted class activation mapping was used to examine the spatial regions contributing to model predictions. As shown in Figure 8, model attention was generally concentrated on lesion regions and diagnostically relevant internal structures. However, some activation maps also extended towards image artefacts and surrounding background areas, indicating that non-lesion visual cues could influence individual predictions.

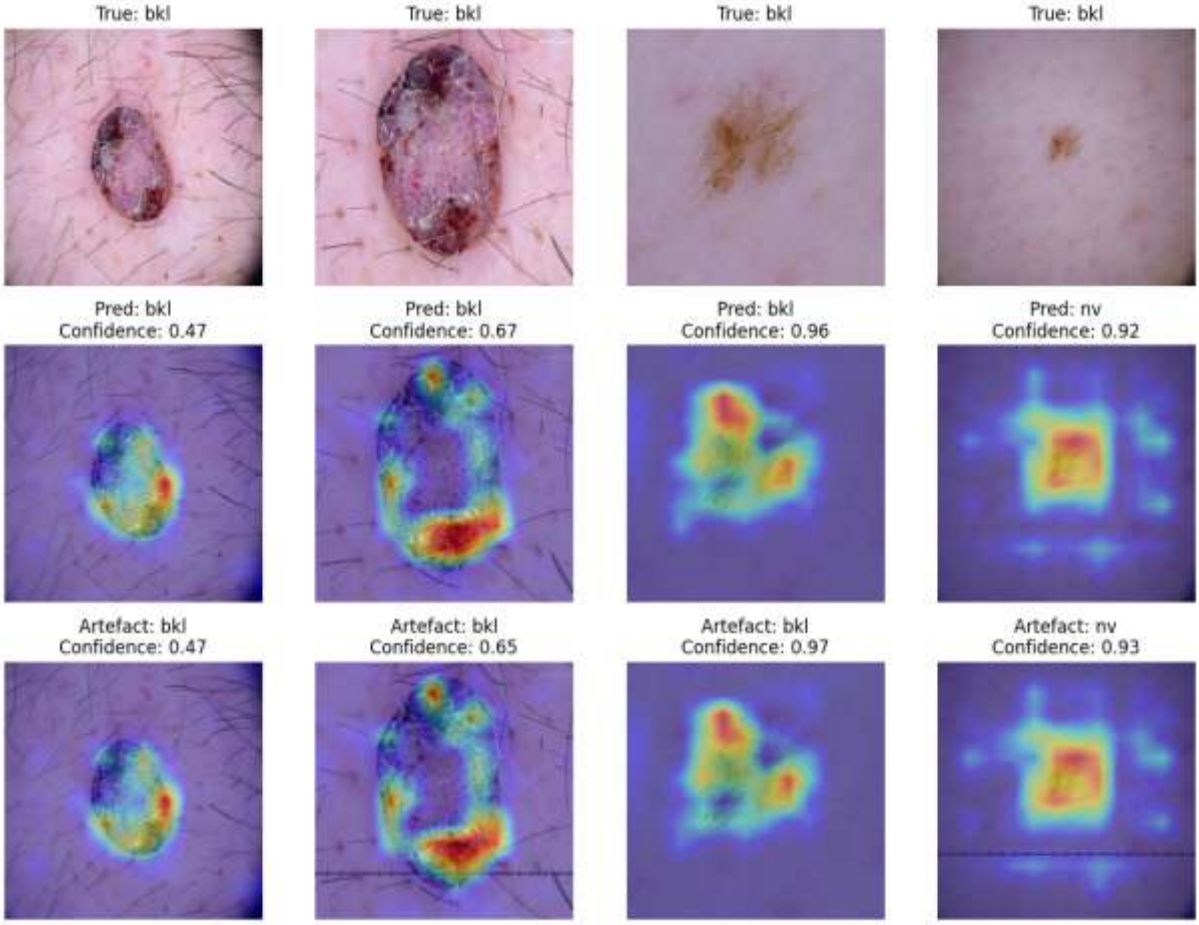


Figure 8. Grad-CAM visualisations of representative benign keratosis-like lesion images, showing the original dermoscopic images, model predictions and confidence scores, and the corresponding attention maps before and after artefact perturbation.

5. DISCUSSION

DermaEdge-X achieved a held-out lesion-disjoint test accuracy of 88.19%, balanced accuracy of 73.99%, macro F1-score of 76.98%, macro ROC-AUC of 0.9695 and macro PR-AUC of 0.8610 under lesion-grouped evaluation. The difference between the weighted F1-score of 88.10% and the macro F1-score reflects the persistent effect of class imbalance, with stronger performance for melanocytic naevi, vascular lesions, basal cell carcinoma and

benign keratosis-like lesions than for melanoma and dermatofibroma. The comparatively low melanoma F1-score of 0.6415 and dermatofibroma recall of 0.3333 indicate that visually overlapping and sparsely represented classes remain difficult to classify reliably. The confusion pattern between melanoma and melanocytic naevi further confirms that high overall discrimination does not eliminate clinically important class-specific errors.

The model showed favourable probability calibration, with an expected calibration error of 0.0379 after temperature scaling, and maintained overall accuracy above 84% under all tested image perturbations. However, balanced accuracy and macro F1-score were more sensitive than overall accuracy to colour, brightness and blur distortions, suggesting that minority-class predictions were particularly affected by image-quality degradation. Computationally, DermaEdge-X required 51.04 GFLOPs and contained 51.19 million parameters, with a median inference latency of 23.85 ms per image on an NVIDIA Tesla T4 GPU. These findings indicate that the framework offers strong discrimination and practical GPU-based inference, although its model size and minority-class sensitivity may limit direct deployment in highly constrained or clinically heterogeneous environments.

6. CONCLUSION AND FUTURE WORK

This study presented DermaEdge-X, a lesion-grouped and class-imbalance-aware framework for seven-class dermoscopic image classification. The model achieved strong discrimination on the held-out lesion-disjoint test set, substantial agreement with reference labels and favourable probability calibration, while retaining relatively stable performance under several controlled image perturbations. The results demonstrate that combining leakage-resistant data partitioning, weighted optimisation, transfer learning and calibration provides a more rigorous assessment than relying on overall accuracy alone.

Future work should prioritise external validation using multi-centre datasets, different imaging devices and broader skin-tone representation. Further investigation is also required to improve melanoma and dermatofibroma recognition through targeted augmentation, cost-sensitive learning, focal or asymmetric losses and uncertainty-aware referral mechanisms. Model compression, pruning, quantisation and knowledge distillation should be explored to reduce storage and computational demands, while prospective clinical evaluation will be necessary to determine the utility of DermaEdge-X as a decision-support tool rather than as a standalone diagnostic system.

Declarations

Ethics statement

This study was conducted using publicly available datasets. No research involving identifiable human participants was carried out. Accordingly, ethical approval was not required.

Consent to participate

Not applicable.

Consent to publish

Not applicable.

Funding

No funding was received for this study.

REFERENCES

[1] Deepa J, “A ROBUST U-NET BASED FRAMEWORK FOR HIGH-FIDELITY SKIN LESION SEGMENTATION,” *ijam*, vol. 39, no. 1s, pp. 321–335, Jan. 2026, doi: 10.12732/ijam.v39i1s.1662.

[2] M. Wang, X. Gao, and L. Zhang, “Recent global patterns in skin cancer incidence, mortality, and prevalence,” *Chinese Medical Journal*, vol. 138, no. 2, pp. 185–192, Jan. 2025, doi: 10.1097/CM9.0000000000003416.

[3] D. S, L. Padma Suresh, and A. John, “A Deep Transfer Learning framework for Multi Class Brain Tumor Classification using MRI,” in *2020 2nd International Conference on Advances in Computing, Communication Control and Networking (ICACCCN)*, Greater Noida, India: IEEE, Dec. 2020, pp. 283–290. doi: 10.1109/ICACCCN51052.2020.9362908.

[4] P. Tschandl, C. Rosendahl, and H. Kittler, “The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions,” *Sci Data*, vol. 5, no. 1, p. 180161, Aug. 2018, doi: 10.1038/sdata.2018.161.

[5] V. D. Nguyen, N. D. Bui, and H. K. Do, “Skin Lesion Classification on Imbalanced Data Using Deep Learning with Soft Attention,” *Sensors*, vol. 22, no. 19, p. 7530, Oct. 2022, doi: 10.3390/s22197530.

[6] T. G. Debelee, “Skin Lesion Classification and Detection Using Machine Learning Techniques: A Systematic Review,” *Diagnostics*, vol. 13, no. 19, p. 3147, Oct. 2023, doi: 10.3390/diagnostics13193147.

[7] H. M. L. Delower *et al.*, “Skin Lesion Detection and Classification Using Machine Learning: A Comprehensive Approach for Accurate Diagnosis and Treatment,” in *2023 26th International Conference on Computer and Information Technology (ICCIT)*, Cox’s Bazar, Bangladesh: IEEE, Dec. 2023, pp. 1–6. doi: 10.1109/ICCIT60459.2023.10441420.

[8] Y. Alkhrijah *et al.*, “From local textures to global attention: a hybrid feature fusion network for skin lesion classification,” *Sci Rep*, vol. 16, no. 1, p. 18609, Apr. 2026, doi: 10.1038/s41598-026-48260-5.

[9] A. S. Al-Waisy, S. Al-Fahdawi, M. I. Khalaf, M. A. Mohammed, B. Al-Attar, and M. N. Al-Andoli, "A deep learning framework for automated early diagnosis and classification of skin cancer lesions in dermoscopy images," *Sci Rep*, vol. 15, no. 1, p. 31234, Aug. 2025, doi: 10.1038/s41598-025-15655-9.
[10] V. Gabani, T. M. Navamani, K. Shyamala, and V. K. Vaswani Rajpal, "Multimodal skin lesion classification for early cancer diagnosis using deep learning," *Front. Physiol.*, vol. 17, p. 1717517, Feb. 2026, doi: 10.3389/fphys.2026.1717517.
[11] I. Aruk, I. Pacal, and A. N. Toprak, "A novel hybrid ConvNeXt-based approach for enhanced skin lesion classification," *Expert Systems with Applications*, vol. 283, p. 127721, Jul. 2025, doi: 10.1016/j.eswa.2025.127721.
[12] Y. Nie, P. Sommella, M. Carratù, M. O'Nils, and J. Lundgren, "A Deep CNN Transformer Hybrid Model for Skin Lesion Classification of Dermoscopic Images Using Focal Loss," *Diagnostics*, vol. 13, no. 1, p. 72, Dec. 2022, doi: 10.3390/diagnostics13010072.