International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 13 Issue: 11 | Nov 2026 www.irjet.net p-ISSN: 2395-0072

AI-Driven Skin Cancer Detection: An Explainable and Reliability-Aware Framework Using EfficientNet-B0

Mr. Giridhar S T[1], Shubhangi[2], Tanushree KA[3], V Ashritha[4]

[1]*Assistant Professor, Department of Computer Science Engineering (Data Science), SJB Institute of Technology, Bengaluru, Karnataka, India*

[234]*Department of Computer Science Engineering (Data Science), SJB Institute of Technology, Bengaluru, Karnataka, India*

Abstract – *Automated recognition of skin lesions and general skin conditions is challenging because of visual similarities between categories, class imbalance, and variation in image acquisition. This paper presents an AI-Driven Skin Cancer Detection application integrating two independent EfficientNet-B0 classifiers. The first is an adopted seven-class dermoscopic lesion classifier based on the HAM10000 dataset of 10,015 original images. Its source-reported test accuracy is 85.77 percent, with macro F1-score 0.7512, expected calibration error 0.0450, and multiclass Brier score 0.2764. It provides class probabilities and Grad-CAM visual explanations; it was not independently trained in this study. The second is a team-trained eight-class general skin-condition model using the Multi-Class Skin Disease Image Dataset with Severity Levels from Mendeley Data. It predicts three acne severity levels, three eczema severity levels, normal skin, and a miscellaneous class. The model used 3,631 training images and 777 validation images over eight epochs, reaching a best validation accuracy of 79.54 percent. This second classifier supports the applications live-camera feature, while the adopted model processes uploaded dermoscopic images. Results are reported separately because the datasets, categories, and evaluation partitions differ. Independent testing, documentation of training settings, and clinical validation remain necessary; this system is intended for educational and research use only.*

Key Words: EfficientNet-B0, HAM10000, skin-lesion classification, general skin conditions, Grad-CAM, probability calibration, live-camera classification, deep learning.

1. INTRODUCTION

The analysis of skin lesions is difficult due to the fact that benign and malignant lesions may have similar colors, textures, shapes, and margins. Dermoscopy affords a magnified view of the skin features and can be used to assist clinical examination, but the specialist examination and, if necessary, histopathological investigation should still be carried out. These challenges motivate research on computer-assisted analysis of dermoscopic images.

Because benign and malignant lesions can share similar visual characteristics, dermoscopy is valuable for examining lesion structures. Nevertheless, clinical assessment and histopathological investigation, when indicated, remain essential. This motivates research into computer-assisted dermoscopic image analysis, where deep learning models have become widely studied classification approaches.

Convolutional Neural Networks (CNNs) are neural networks that automatically learn visual features from annotated data. Esteva et al. demonstrated the potential of deep neural networks for skin cancer image classification [1]. However, high accuracy on one benchmark does not establish generalization across all patient populations and imaging conditions.

Several factors affect the interpretation of classification results, including class imbalance, dataset diversity, possible overlap between images of the same lesion, evaluation metrics, and prediction confidence. Overall accuracy may be high even when clinically important melanoma cases are missed. This paper presents a narrative review and an educational case study integrating an adopted seven-class EfficientNet-B0 dermoscopic classifier with a separately team-trained eight-class skin-condition classifier. Performance and implementation details of the seven-class model are attributed to the original project documentation [2].

The main contribution is an application that integrates two independent classification pathways: one for uploaded dermoscopic skin-lesion images and another for live-camera images of general skin conditions. The second pathway distinguishes acne and eczema severity levels, normal skin, and a miscellaneous category. The two classifiers retain separate datasets and evaluation procedures.

2. REVIEW SCOPE AND APPROACH

This paper adopts a narrative review approach. A selection of literature was used to provide an overview of the classification of skin lesions, dermoscopic data sets, CNN architectures, transfer learning, and interpretability

and evaluation. This paper draws on foundational literature, benchmark information, and recent reviews [3, 4, 5]. It is a narrative review rather than a systematic review; no PRISMA-style screening or exhaustive literature search is claimed.

The discussion also includes source-reported results from the adopted educational project and its repository [2]. These are discussed as part of a case-study. As the literature does not all use the same data sources, definitions of classes, training budget, image processing, data splitting and testing protocol, there is no way to compare results between studies.

3. DATASETS FOR SKIN-LESION AND GENERAL SKIN-CONDITION CLASSIFICATION

3.1 HAM10000 Dataset

HAM10000 is a publicly available collection of 10,015 dermoscopic images of common pigmented skin lesions acquired from multiple sources [3]. It contains seven diagnostic categories: actinic keratoses and intraepithelial carcinoma (akiec), basal cell carcinoma (bcc), benign keratosis-like lesions (bkl), dermatofibroma (df), melanoma (mel), melanocytic nevi (nv), and vascular lesions (vasc). Associated metadata may include age, sex, anatomical location, and lesion identifiers, which can support subgroup analysis and lesion-aware data partitioning when available.

A central challenge is the strong imbalance between diagnostic categories. Melanocytic nevi constitute the majority of the dataset, whereas dermatofibroma and vascular lesions have relatively few examples. Consequently, high overall accuracy may conceal weaker performance on minority classes. Class-wise precision, recall, and F1-scores are therefore essential for evaluating the source-reported seven-class model. Dataset composition alone does not guarantee representation of all clinical populations, skin tones, or image-acquisition conditions.

Table -1: Class Distribution of the HAM10000 Dataset

Class	Number of Images
Actinic Keratoses (akiec)	327
Basal Cell Carcinoma (bcc)	514
Benign Keratosis (bkl)	1,099
Dermatofibroma (df)	115
Melanoma (mel)	1,113
Melanocytic Nevi (nv)	6,705
Vascular Lesions (vasc)	142
Total	10,015

3.2 General Skin-Condition Dataset

In addition to the HAM10000 dataset, we have also used the Multi-Class Skin Disease Image Dataset with Severity Levels, obtained from Mendeley Data [6]. This dataset contains images of various skin diseases with Severity Levels. It was used for a separate classification component.

We used this dataset to train an 8 class classifier for general skin diseases. This included mild acne, moderate acne, severe acne, mild eczema, moderate eczema, severe eczema, normal skin and a miscellaneous class called Xyz. This allows us to classify the various skin diseases based on severity level.

The eight-class classifier was trained to support the live-camera classification feature. It was trained using 3,631 images. We used 777 images for validation. In total there were 4,408 images. The model was trained for 8 epochs and had a best validation accuracy of 79.54%.

This dataset is slightly different to the HAM10000 dataset as it deals with general skin diseases rather than dermoscopic skin lesions. This means that the 7 class model is not trained by us but taken from the G26karthik/Skin-Disease GitHub page. Whereas the 8 class model has been trained by us. As they are different datasets and classifying different things the models are not trained together.

3.3 Class Imbalance and Data Splitting

HAM10000 is highly imbalanced. There are more melanocytic nevi than other classes such as dermatofibroma and vascular lesions. This means that a classifier can have high overall accuracy by simply learning to predict the most common classes well, but will perform badly on the less common ones. Therefore, it was important to report the macro F1, precision and recall for each class and the confusion matrix.The images in HAM10000 may be highly correlated if they come from the same lesion. If correlated images are found in both the training and testing subsets then this may give an over-optimistic estimate of how well the model will generalise to completely unseen lesions.

The project summary describes how the images were split into lesion based subsets to mitigate this. Additionally, an experiment should be reproducible so the split, random seed, number of examples per class and any patient/lesion IDs should be recorded. To ensure the model is robust to different acquisition devices, centres, skin tones and patients, independent datasets are required. The International Skin Imaging Collaboration have provided valuable public benchmark datasets, however care must be taken when comparing performance between different datasets [7].

The class distribution shows a substantial imbalance,with melanocytic nevi representing the majority of the original dataset. Therefore, the evaluation considers class-wise precision, recall, and F1-score alongside overall accuracy to better characterize performance across the seven lesion categories.

The project documentation describes lesion-level data partitioning to reduce the possibility of related lesion images appearing in different evaluation subsets. However, the precise numbers of images assigned to training, validation, and testing must be confirmed using the original split records. The split proportions, random seed, and class-wise partition sizes should be reported after verification. These details are important for assessing reproducibility and ensuring that the evaluation reflects performance on previously unseen lesions.

4. DEEP-LEARNING ARCHITECTURE

4.1 Convolutional Neural Networks

CNNs apply a succession of convolutions and non-linearities to an image. Typically early layers will extract basic features such as edges, textural and colour changes. Deeper layers will extract higher level patterns. The network could also extract incidental correlations from the image capture or background.Successful performance on held-out images does not necessarily indicate that the network has learned clinically relevant lesion characteristics, as predictions may also be influenced by background patterns or imaging artefacts.

4.2 Residual Networks

Residual networks such as ResNet [8] take advantage of shortcut connections to help optimize the deep networks. These are commonly used as baseline models when performing image classification research. The usefulness of ResNet to skin lesions will depend on the amount of data available, computational resources, and how they are assessed. To fully compare ResNet with other models it needs to be tested under the same data splits, preprocessing steps, optimization resources and reporting. No novel controlled baseline is created in this case study.

4.3 EfficientNet

EfficientNet employs compound scaling to ensure model depth, width and resolution are in sync [9]. EfficientNet-B0 in particular is a reasonably small model that provides an ideal balance between computational cost and image classification ability. The adopted seven-class model uses an ImageNet-pretrained EfficientNet-B0 backbone and a task-specific classification head that produces seven scores. The model has been chosen based on a reasonable cost to train and inference to allow it to be used for learning purposes. This does not mean it is the best option for the problem, there is no evidence to suggest it will perform better than any alternative.

4.4 Transfer Learning and Augmentation

Transfer learning is used so that the representation created from the large dataset can be adapted to the new task. The weights of the pre-trained ImageNet model are used as an initialisation for the dermoscopic image classification task. The final output layer has been created which will map the features to the seven classes of skin lesions.

The adopted project describes basic training-time image augmentation, including rotation, flipping, and changes in colour or brightness. Random augmentation should be confined to training data; validation and test images require deterministic preprocessing. More extensive augmentation using Albumentations is a possible future extension, not a verified part of the eight-class training procedure.

5. PREPROCESSING AND EVALUATION

5.1 Image Preparation

The dermoscopy image is first converted into a uniform colour representation and resized to have dimensions of 224×224 pixels. This is then converted into a tensor representation and normalised according to the channel statistics of ImageNet. This ensures that the data matches what would be expected when using weights which have been pre-trained.The model outputs 7 un-normalised scores known as logits.

Using a softmax layer these can be converted into probabilities that are positive and sum to 1:

$$p_c = \frac{\exp(z_c)}{\sum_{j=1}^{7} \exp(z_j)}, \qquad c = 1, \ldots, 7 \qquad (1)$$

Where z_c is the logit output from the model for class c. The category with the highest probability will give the predicted category. However the full probability vector offers a more informative way of understanding how confident the model is about its prediction.

5.2 Classification Metrics

Accuracy represents the fraction of correctly predicted test cases:

$$\text{Accuracy} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}\,(\hat{y}_i = y_i), \qquad (2)$$

where N is the number of test images, y_i is the reference label, and $\hat{y}_i$ is the predicted label.

For each diagnostic class c, precision and recall are

$$P_c = \frac{TP_c}{TP_c+FP_c}, \qquad R_c = \frac{TP_c}{TP_c+FN_c}, \tag{3}$$

where TP_c, FP_c, and FN_c denote true positives, false positives, and false negatives for that class.

The F1-score is the harmonic mean of precision and recall:

$$F1_c = \frac{2P_cR_c}{P_c+R_c}. \tag{4}$$

The macro F1-score is the average of class-wise F1-scores. It equally weights the 7 classes in the dataset. This gives insight into how well the model is performing on an imbalanced lesion dataset, which may be hidden by looking at the overall accuracy.

5.3 Probability Calibration and Reliability

Besides classification accuracy, the new project considers calibration of model predicted confidence. Expected calibration error (ECE) takes predictions and sorts them into confidence groups (or bins). The error is calculated by looking at the difference between confidence and accuracy within each group. [10]. For M bins B_m,

$$ECE = \sum_{m=1}^{M} \frac{|B_m|}{N} |acc(B_m) - conf(B_m)|. \tag{5}$$

The multi-class Brier score measures mean squared difference between predicted probabilities and one-hot reference labels:

$$Brier = \frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{7}(p_{ic} - y_{ic})^2. \tag{6}$$

A reliability diagram plots empirical accuracy against estimated confidence to help identify overconfidence or underconfidence.

For these metrics, lower values generally indicate better agreement or lower squared probabilistic error, respectively, under the stated conventions. These are assessment techniques. Calculating the ECE and Brier scores do not necessarily lead to well-calibrated predictions. The summary given is for assessment of raw softmax outputs. The use of temperature scaling and other post-hoc calibration techniques will be an area of future research. However, calibration will not necessarily generalise across populations.

5.4 Inference Performance Profiling

The project also includes average inference profiling. Latency is the time taken to perform an inference and throughput is how many images are processed per second. These were measured on a prototype using GPU so cannot be considered as validated end to end processing times. The runtime will depend heavily on the hardware used, the batch size, image loading times, data transfer, warmup and how you measured the runtime. The experiment could be repeated on more devices and include some variation. It would also be useful to separate the runtimes into preprocessing, network inference, and heatmap creation.

6. EXPLAINABLE ARTIFICIAL INTELLIGENCE

Gradients are used to produce a weighted map of feature maps in a convoluted layer for a specified class (Grad-CAM) [11]. The prototype uses a bespoke version of Grad-CAM to provide a heat map overlay of the final convolutional features. This allows the interface to display both the predicted class probabilities and the visual explanation.

This does not necessarily mean the area is showing any diagnostic value. A heat map may be produced that looks realistic even if the classification was incorrect. Also, the heat map may just be highlighting background artefacts rather than useful features of the lesion. To assess how well the heat map is explaining the model, other methods should be used. This includes comparing them to expert lesion annotations, conducting perturbation analysis and performing usability tests with clinicians. Other methods such as SHAP, LIME or counterfactual explanations can be considered as future research but are not confirmed to be implemented based on the project summary provided.

6.1 Grad-CAM Visualization Examples

Grad-CAM visualizations provide a qualitative method for examining the image regions associated with model predictions. The generated heatmaps can be overlaid on the corresponding dermoscopic images to support visual interpretation.

These explanations should be considered alongside the predicted classification and reference label. Highlighted regions do not independently demonstrate medical correctness or validated lesion localization.

7. COMPARISON OF METHOD FAMILIES

Different approaches to feature engineering, CNN training from scratch, pre-training, ensemble methods, and explainability have been explored [4, 5]. Vision transformers provide another image-classification architecture family [12], although no transformer baseline

is evaluated here. These methods vary in their data, computational and interpretability requirements.

The following provides some general observations on the strengths and weaknesses of each method. This is not a benchmarking study, as the methods have not been applied with a unified experimental protocol.

Table -2: Qualitative Comparison of Skin-Lesion Methods

Approach	Strength and principal limitation
Handcrafted features + ML	Lightweight; depends strongly on feature design and image quality.
CNN from scratch	Learns task-specific features; requires adequate labelled data and regularization.
ImageNet transfer learning	Provides reusable visual features; outcome depends on fine-tuning and data alignment.
EfficientNet family	Balances scale and compute; comparisons still require matched experimental conditions.
Ensembles and hybrids	May combine complementary predictions; increase complexity and inference cost.
Grad-CAM and related XAI	Visualize model attention; do not establish medical correctness or explanation fidelity.

8. DUAL-MODEL SYSTEM IMPLEMENTATION AND EVALUATION

8.1 System Overview

This AI Driven Skin Cancer Detection application combines two separate deep learning models to help classify skin images in different ways. The main aim of this is to allow the user to utilise two different classifications, one for dermoscopic skin lesion detection and another for general skin condition detection. Both models are using the EfficientNet-B0 architecture. However, they have been trained differently for different purposes.

The first model is a pre-trained 7 class skin lesion classifier taken from the G26karthik/Skin-Disease GitHub page [2]. This model uses the EfficientNet-B0 architecture with PyTorch. This model was trained using the HAM10000 data set of dermoscopic skin lesions. This model was adopted from the GitHub project and was not independently trained or fine-tuned in the present study.

The model takes the uploaded dermoscopic image and predicts what class the skin lesion belongs to, with 7 different classes to choose from. The model will predict the class with a probability score for each class. This will show the user what class it thinks the image belongs to and how confident it is with each class. The Grad-CAM feature is used with this model to show what areas of the image it is looking at. This gives the user extra information about how the model made its prediction. However, the areas shown do not necessarily indicate any important medical features of the skin lesion. This means the prediction does not indicate if it is correct.

The second model is an 8 class skin condition classifier. This model was trained separately by the project team using the Multi-Class Skin Disease Image Dataset with Severity Levels from Mendeley Data [6]. This also uses the EfficientNet-B0 architecture with PyTorch. This was trained separately to the 7 class model. This model has 8 classes it can predict, these are mild acne, moderate acne, severe acne, mild eczema, moderate eczema, severe eczema, normal skin and Xyz (miscellaneous).

The eight-class model used 3,631 training images and 777 validation images. This resulted in 4408 images being used in total. It was trained for 8 epochs and the best validation accuracy achieved was 79.54%. This model takes images of size 224×224 and uses the normalisation values in the saved checkpoint. This indicates the models performance on the validation set and does not indicate how well it performed on a test set. Testing would need to be done on unseen data.

This application has a live camera feature for the 8 class skin condition classifier. This allows the user to take a picture of their skin and have it classified into one of the general skin conditions. The image is then run through the separate trained model. The 7 class EfficientNet-B0 model was taken from the GitHub page and is meant to be used for classifying uploaded dermoscopic images.

Both models are completely separate. The 7 class model taken from the GitHub page has a test accuracy of 85.77%. The 8 class model achieved a best validation accuracy of 79.54%. These two accuracies cannot be compared as they are completely different tasks and were tested on different data sets. This demonstrates the integration of multiple skin-image classification features within one research application. However further testing would need to be done independently. The models would need to be tested on different images and validated by a medical professional before they could be considered suitable for medical decisions. This application is purely for educational/research purposes and should not be used as a substitute for a dermatologist.

8.2 Training Setup and Implementation

There are two separate classifiers used in this system. The first is an existing 7 class skin lesion classifier. The second is a new 8 class general skin condition classifier.

The 7 class skin lesion classifier uses the EfficientNet-B0 architecture, implemented via PyTorch. This model uses ImageNet pre-trained weights and takes an input size of 224 × 224 pixels. The batch size and mixed precision training were also provided by the source repository. [2] The model was not trained or fine-tuned for this study.

HAM10000 contains 10,015 original dermoscopic images across seven categories [3]. The adopted project documentation separately reports 17,036 training-stage image records [2]. The reason for this larger figure has not been independently verified; therefore, it must not be interpreted as a confirmed count of unique HAM10000 images or a verified training-set size.

The EfficientNet-B0 network has been used for the 7 class skin lesion classification. It has already been trained so we have used these weights to make predictions. This network has not been re-trained or fine-tuned by us. We have used the network from the original repository with the training settings and results. We have simply used the classifier in our application.

The second classification model has been trained separately to provide further functionality. The Multi-Class Skin Disease Image Dataset with Severity Levels has been obtained from Mendeley Data [6]. There are 8 classes in this data set; mild acne, moderate acne, severe acne, mild eczema, moderate eczema, severe eczema, normal skin and a miscellaneous category called Xyz. These images are a general skin condition classifier.

A total of 4,408 images were used across the training and validation partitions: 3,631 for training and 777 for validation. The eight-class model uses EfficientNet-B0 in PyTorch with input images of 224 × 224 pixels. Its saved checkpoint specifies normalization means of [0.485, 0.456, 0.406] and standard deviations of [0.229, 0.224, 0.225]. After eight reported training epochs, the best recorded validation accuracy was 79.54%.

This 79.54% result measures performance on the validation partition, not on an independent test dataset. External testing is still needed to evaluate generalization to unseen skin images.

The 8 class network provides the live camera functionality in the expanded application. The images captured by the camera will be inputted into the general skin condition classification. The model has been trained and will make a prediction about which of the 8 classes the skin image belongs to.

The two networks have different data and outputs. The 7 class network has been used for dermoscopic lesion classification. The 8 class network has been trained by us for general skin condition detection. The networks have been trained differently as the 7 class model was taken from an existing network. The accuracies should not be compared as they are different.

The eight-class classifier is based on an EfficientNet-B0 architecture. It was trained using PyTorch and the saved checkpoint shows that it has eight output classes, an image size of 224x224 pixels, and ImageNet normalization parameters. The original eight-class training script was unavailable for verification of the optimizer, learning rate, batch size, loss function, random seed, and checkpoint-selection method. Including this information will improve reproducibility of the results and will help when comparing experiments in the future.

Each network will be shown separately. The 7 class network will show the results from the source. The 8 class network will show the accuracy from the validation. The networks are available within the same application but each have their own claims about how well they perform.

Further testing of the 8 class network on an external dataset would be useful to see how well it can generalise. This would include different lighting conditions, cameras, backgrounds and skin types. This is a key point for live camera applications as the images may look very different to those used in the training. This has been implemented to show how two classification systems can be combined for educational and research purposes. However, more validation would be required for the models to make the predictions suitable for decision making in a clinical setting.

8.3 Evaluation Results of the Two Models

The two classifiers have different evaluation histories, and their results are reported separately. The 7 classes EfficientNet-B0 classifier has been reported to have a test accuracy of 85.77%, with a macro F1-score of 0.7512 [2]. It has also been reported that the model has an ECE of 0.0450, as well as a multi-class Brier score of 0.2764. On average, it takes 6.29ms for the model to make predictions on the GPU and can therefore process images at a rate of 158.89 images/s.

The 7 class metrics reported above are based on the original GitHub project. The 8 class general skin condition classifier has been trained separately as part of this project from 3,631 training images and 777 validation images. It has been trained for 8 epochs.

The best validation accuracy for the 8 class classifier was 79.54%, which was confirmed by the model checkpoint file.

As the two models have been trained on different data sets, with different categories and methods of evaluation

the accuracy values cannot be compared. Further independent testing of the 8 class classifier will be needed to determine how well the model has generalised to unseen skin images.

Table -3: Reported Evaluation Outcomes of the Two Classification Models

Metric	Value
Model 1: Adopted Seven-Class Model	
Source-reported test accuracy	85.77%
Source-reported macro F1	0.7512
Expected calibration error	0.0450
Multiclass Brier score	0.2764
Mean GPU inference latency	6.29 ms/image
Reported throughput	158.89 images/s
Model 2: Team-Trained Eight-Class Model	
Training images	3,631
Validation images	777
Completed training	8 epochs
Best validation accuracy	79.54%

8.4 Training and Validation Performance

Training and validation curves can help evaluate model performance across successive epochs. However, complete epoch-wise training records were not available for inclusion in this study. Therefore, only the best validation accuracy of 79.54% is reported for the eight-class classifier. No independent eight-class test accuracy or epoch-wise learning curves are claimed.

If authenticated epoch-wise records become available, their trajectories could be examined to assess convergence and differences between training and validation performance.

8.5 Per-Class Performance and Melanoma Detection

This table displays accuracy information including precision, recall and F1-scores for each of the 7 classes. We can see that the recall for melanocytic nevi is much higher compared to melanoma (0.9450 vs 0.6310). This demonstrates the importance of using macro-averaging and per-class evaluation on imbalanced data. The reported melanoma recall of 0.6310 indicates that 63.10% of melanoma test images were correctly identified, while approximately 36.90% were incorrectly assigned to other lesion categories. This finding highlights an important limitation that must be considered when interpreting the overall classification accuracy. This is a significant issue when considering any proposed clinical triage application. The precision of 1.0000 for vascular lesions shouldn't be taken as perfect. This is only the precision achieved during testing. The complete support and confidence intervals for each class would need to be known in order to understand the statistical uncertainty.

Table -4: Reported Per-Class Test Performance

Class	Precision	Recall	F1	Support
nv	0.9166	0.9450	0.9306	2000
mel	0.7067	0.6310	0.6667	336
bkl	0.7467	0.7134	0.7296	314
bcc	0.7534	0.7143	0.7333	154
akiec	0.6667	0.6275	0.6465	102
vasc	1.0000	0.7727	0.8718	44
df	0.6071	0.7727	0.6800	44

8.6 Confusion Matrix Analysis

A seven-class confusion matrix provides a detailed representation of the relationship between the reference lesion categories and the predictions generated by EfficientNet-B0. The diagonal entries indicate correct classifications, whereas off-diagonal entries represent misclassified samples.

The confusion matrix enables identification of frequently confused lesion categories and provides additional insight into class-specific classification errors.

8.7 ROC and Precision–Recall Analysis

Receiver operating characteristic (ROC) analysis evaluates the ability of a classification model to distinguish individual lesion categories from the remaining classes at different thresholds.

For the seven-class classification task, a one-versus-rest strategy can be used to calculate the ROC curve and area under the curve (AUC) for each diagnostic class. Precision–recall analysis provides additional information about model performance on imbalanced datasets. It describes the relationship between precision and sensitivity across different classification thresholds.

This analysis is particularly relevant for evaluating minority classes including melanoma, dermatofibroma, and vascular lesions. ROC and precision–recall curves are discussed as evaluation methods; however, independently verified curve results are not reported in the present study.

8.8 Calibration, Interpretability, and Speed

Calibration is an important factor in assessing how well a deep learning model performs. In a medical image classification task it is important to understand how confident the model is with its predictions. A model could have a high classification accuracy but assign very high confidence to incorrect predictions. The seven class EfficientNet-B0 model used has an expected calibration

error (ECE) of 0.0450. This shows there is a small difference on average between the confidence of the predictions and the actual accuracy for the given binning method. The multiclass Brier score is 0.2764. This gives the squared error of the predicted probabilities compared to actual class labels. This can be visualised using a reliability diagram. However, this does not confirm if the model is reliable in a clinical setting [2].

The Grad-CAM visualisations show further details of the image areas which are being used to make classifications. The heatmap can help with qualitatively interpreting predictions, but these do not provide medical accuracy in identifying lesions within the image. Grad-CAM and probability calibration assess different aspects of model behaviour.

The model has a mean latency of 6.29 ms per image using a GPU according to the source. The source also reports a throughput of 158.89 images/s [2]. These source-reported measurements describe the original prototype and do not represent independently verified end-to-end latency of the integrated application. Performance can vary depending on the hardware used, image pre-processing and other factors. Further evaluation is needed to determine how the model would perform in practice.

8.9 Reproducibility Considerations

Reproducible experiments should document data manifests, dataset provenance, augmentation, random seeds, library versions, checkpoint-selection criteria, and class-wise evaluation results. The adopted repository describes lesion-aware splitting, but exact split records and the basis of its 17,036 training-stage image records have not been independently verified. In the present study, the team-trained eight-class model has a recorded validation result but no independently reported external test result. These limitations are explicitly acknowledged rather than presented as verified experimental details.

By continually adjusting the model choices based on the held out test results optimistic estimates can be obtained. Future studies should ensure a final untouched test set is used to separate hyperparameter selection, calibration fitting and final evaluation.

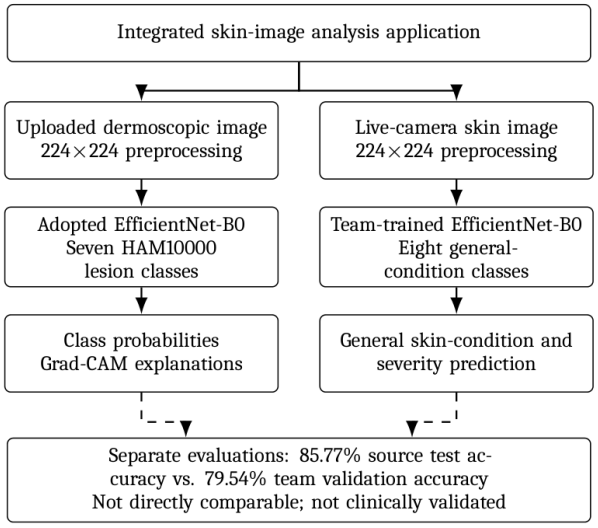


Fig -1: Dual-model application workflow. Dermoscopic uploads use the adopted seven-class classifier; live-camera images use the team-trained eight-class classifier. Their metrics come from different evaluation settings.

9. LIMITATIONS AND RESEARCH GAPS

9.1 Class Imbalance and Melanoma Sensitivity

The data is skewed across the 7 classes and so the under-represented classes may be challenging to correctly classify. Whilst the macro F1-score is informative, it does not solve the underlying issue of data imbalance and does not guarantee good performance on the minority classes. The melanoma recall is lower than the target of 0.85 sensitivity in the project summary. Future work could explore using class weighted or focal loss functions, along with a balanced sampling scheme, data augmentation and threshold selection [13]. Any changes should be tested on unseen data.

9.2 Generalization, Fairness, and Representation

The evaluation is limited in that it does not cover performance across different imaging devices, clinical centres, patient populations and skin tones. Public dermoscopic data sets may lack representation of certain demographic and images may not be representative of those captured in real world scenarios. External validation and subgroup analysis could allow for insights to be gained on how models perform based on available attributes (age, sex, location and skin tone), providing sufficient sample sizes and ethical data handling. However the project summary provided does not state that any subgroup fairness results have been obtained.

9.3 Calibration and Clinical Utility

The ECE and Brier scores provide additional information on how good the predicted probabilities are. However, they are sensitive to the distribution of evaluations and definition of metrics. The project currently uses raw probabilities from the softmax function rather than a validated temperature scaled model. The target ECE score, of less than 0.035, has not been achieved (score = 0.0450).

In practice, this model would need to be deployed with appropriate thresholds, sensitivity analysis, clinician review, prospective validation, data protection and regulatory review. This is not demonstrated in this example.

9.4 Explainability and Uncertainty

This method can be used as a baseline to visually inspect the model, however it is not validated as a causal explanation. Future experiments could compare the heatmaps against expert annotations. The model could be tested for robustness under slight changes in brightness, contrast or rotation.

This summary has not detailed advanced uncertainty methods such as Monte Carlo dropout or deep ensembles. The softmax confidence value is not a full uncertainty measure.

Table -5: Reported Goals Versus Observed Prototype Results

Measure	Reported	Stated goal
Macro F1	0.7512	≥ 0.78; not met
Melanoma recall	0.6310	≥ 0.85; not met
ECE	0.0450	< 0.035; not met
Mean GPU latency	6.29 ms	< 8 ms; met in prototype
Throughput	158.89 img/s	≥ 150 img/s; met in prototype

10. CONCLUSION

The AI-Driven Skin Cancer Detection application integrates two distinct EfficientNet-B0 classification pipelines. The first is a seven-class dermoscopic skin-lesion model adopted from the G26karthik/Skin-Disease repository and originally developed using HAM10000. It provides lesion-category probabilities and Grad-CAM explanations. The reported test accuracy of 85.77% and macro F1-score of 0.7512 originate from the source project and do not represent independent training or evaluation conducted in this study [2].

The second pipeline is an eight-class general skin-condition model trained separately by the project team on the Mendeley dataset. It supports live-camera classification of acne and eczema severity, normal skin, and miscellaneous conditions. Training used 3,631 images, with 777 validation images, over eight epochs; the best recorded validation accuracy was 79.54%. This validation metric is not directly comparable with the first model's source-reported test accuracy. The combined application demonstrates two functional analysis pathways, but incomplete training documentation, the lack of an independent eight-class test set, and the need for clinical validation limit its use to education and research rather than medical diagnosis.

REFERENCES

[1] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun, "Dermatologist-level classification of skin cancer with deep neural networks," Nature, vol. 542, no. 7639, pp. 115–118, 2017, doi: 10.1038/nature21056.

[2] G26karthik, "Skin-Disease: AI-assisted dermatological decision-support system," GitHub repository. [Online]. Available: https://github.com/G26karthik/Skin-Disease. Project metrics and configuration are also reported in the supplied SUMMARY_PBL.md.

[3] 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, Art. no. 180161, 2018, doi: 10.1038/sdata.2018.161.

[4] T. G. Debelee, "Skin lesion classification and detection using machine learning techniques: A systematic review," Diagnostics, vol. 13, no. 19, Art. no. 3147, 2023, doi: 10.3390/diagnostics13193147.

[5] P. Hermosilla et al., "Skin cancer detection and classification using neural network algorithms: A systematic review," Diagnostics, vol. 14, no. 4, Art. no. 454, 2024, doi: 10.3390/diagnostics14040454.

[6] A. M. Ansari, L. Shakeel, S. M. B. Qasim, and M. Parveen, "Multi-Class Skin Disease Image Dataset with Severity Levels," Mendeley Data, 2026. [Online]. Available: https://data.mendeley.com/datasets/6sjcrtg54p.

[7] N. C. F. Codella et al., "Skin lesion analysis toward melanoma detection: A challenge at the 2017 International Symposium on Biomedical Imaging," in Proc. IEEE 15th Int. Symp. Biomedical Imaging, pp. 168–172, 2018, doi: 10.1109/ISBI.2018.8363547.

[8] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proc. IEEE Conf. Computer Vision and Pattern Recognition, pp. 770–778, 2016, doi: 10.1109/CVPR.2016.90.

[9] M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in Proc. 36th Int. Conf. Machine Learning, PMLR, vol. 97, pp. 6105–6114, 2019.

[10] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in Proc. 34th Int. Conf. Machine Learning, PMLR, vol. 70, pp. 1321–1330, 2017.

[11] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in Proc. IEEE Int. Conf. Computer Vision, pp. 618–626, 2017, doi: 10.1109/ICCV.2017.74.

[12] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in Proc. Int. Conf. Learning Representations, 2021.

[13] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in Proc. IEEE Int. Conf. Computer Vision, pp. 2980–2988, 2017.