RESEARCH PAPER

A Multi-Scale Cross-Scale Fusion Network for Explainable Multi-Class Skin Disease Classification

Tushar Surwadkar, Dr. Sujata Deshmukh

1Research Scholar, Fr. Conceicao Rodrigues College of Engineering /Computer Engineering Department, Mumbai, India
Email: tusharsphd@gmail.com, https://orcid.org/0009-0006-3047-0603
2Professor, HoD, Fr. Conceicao Rodrigues College of Engineering / Computer Engineering Department, Mumbai, India
Email: sujata.deshmukh@fragnel.edu.in, https://orcid.org/0000-0001-9893-6947

Abstract

It is difficult to classify skin diseases based on medical images automatically - although the pattern for diagnosis is the same visually, it hides itself at different spatial scales. We establish and test a multi-scale convolutional neural network with the cross-scale feature fusion for five types of skin disease classifications - nail psoriasis, SJS-TEN, Vitiligo, acne, and hyperpigmentation. We assessed five models in the process: M0 is a standard convolutional neural network, M1 introduces a multi-scale feature extraction process, M2 involves the process of cross-scale feature fusion, M3 adds the convolutional block attention (CBAM), and lastly, M4 has batch normalization in the classifier head. The feature extraction process consists of using parallel convolutional methods with 3×3, 5×5, and dilated 3×3 convolutions plus the principle of residual learning. Regarding the evaluation of M3 on the independent test dataset of 1,426 images, the system obtained the highest performance results equal to 95.86% accuracy, 96.93% balanced accuracy, and 95.13% Macro-F1 score. It needs to be noted that M3 got macro-Precision of 93.77% and specificity of 99.01%. The McNemar test for the analysis of the outcomes obtained using M2 and M3 showed the difference to be significant ($p = 0.00352$). Moreover, calibration analysis was carried out along with the application of Grad-CAM. In addition, the complete duplication control confirmed that the dataset included similar images in different parts of the gathered data; thus, the reported results should be treated with caution.

Keywords: Skin disease classification; deep learning; convolutional neural networks; multi-scale feature extraction; cross-scale feature fusion; CBAM; explainable artificial intelligence; Grad-CAM; medical image analysis; dermatological image classification.

How to cite this article: Tushar Surwadkar Dr. Sujata Deshmukh. A Multi-Scale Cross-Scale Fusion Network for Explainable Multi-Class Skin Disease Classification. Int J Drug Deliv Technol. 2026;16(82s):459-469. DOI: 10.25258/ijddt.16.82s.55

Introduction

Skin diseases are highly varied in their form and appearance, colour and texture, lesions and the areas they affect. This makes skin image classification a complicated issue in computer vision because important clinical information can present itself in various ways. For example, the fine texture of parts of the skin can contribute to image classification, but in addition to details of this type, the significance of structure and other general aspects should not be disregarded. Deep convolutional neural networks (CNNs) are extensively applied in medical image recognition as deep learning methods allow for the creation of informative feature representations from low-level and high-level image features. At the same time, traditional sequential CNN structures are unlikely to provide good results in feature representations. This is particularly important in the case of skin images, where diagnostic signs may be found in both fine texture and big parts of the image. This research article discusses the development of a research survey to investigate the subject in question. The process of this experimentation progresses from a classic CNN model to the creation of different multi-scale and cross-scale fusion systems. Instead of employing different modifications of designs at once, the technique allows assessing their performance one by one. Five models are considered: M0 as a basic CNN architecture; M1 as a stochastic model based on the idea of multi-scale feature extraction; M2 based on the cross-scale fusion evidence; M3 employing a convolutional block attention module (CBAM); M4 representing modifications of the classifier head. The multi-scale feature block consists of parallel computation lines that include Convolutional processes with 3×3, 5×5, and dilated 3×3 kernels allowing to receive different outputs that will then be merged with the help of 1×1 convolutions while achieving success in application of another method of data transformation – residual connection. Batch normalization and SiLU activation were applied in the extraction blocks.

A second aspect of the proposed framework involves utilizing features across different depths of networks. The three depth features are merged into a unified 256-channel representation before

combining them into a single representation. Then the 768-channel output is compressed into 256 channels with a 1×1 convolution to allow features from shallow, medium, and deep levels to form implications for the representation. The authors compared five modifications of the model in terms of various classification performance metrics such as accuracy, balanced accuracy, macro-averaging metrics, sensitivity, and specificity. Moreover, the authors used a number of metrics in addition to the confusion matrix and class-wise analysis.

2. Related Work

2.1 Deep Learning for the Classification of Skin Diseases

Using deep learning for skin disease classification has resulted in much progress made in the automatic processing of images in dermatology due to its ability to differentiate diseases relying on visual features. For example, the research conducted by Esteva et al. showed how deep neural networks can classify skin cancer with results like those made by dermatologists, which is a very important achievement in the field of computer-based diagnosis in dermatology [1]. The studies conducted later revealed that deep learning, especially CNN-based methods, is widely used for skin lesions classification, but there are still problems with class imbalance, datasets, robustness, and cross-domain generalization [2,3]. Several further systematic reviews have mapped this rapidly growing field, synthesizing methodological trends, dataset characteristics, segmentation and classification pipelines, and open challenges across the wider literature [8,9,10,11,12,13].

2.2 Multi-Scale Feature Learning

Dermatological images incorporate spatial information in various large directions; fine characteristics such as textures, colour variations, and small lesions may contain useful information for diagnostic purposes, while large-scale structures may describe an overall appearance of the condition. Consequently, multi-scale learning proved to be valuable for receiving complementing information from various receptive fields.

In CNN-based structures, multi-scale representation is achieved through various convolutional functions. In this work, the multi-scale feature extraction module has three parallel branches with kernels of 3×3, 5×5, and dilated 3×3. Such branches possess different receptive fields but have a common feature representation. Thus, after mixing their outputs, the informativeness of features is significantly increased. Related multi-scale attention-based CNN designs have similarly been proposed for skin lesion classification, combining grouped multi-scale attention with class-specific loss weighting to address category imbalance [4].

Using different scales appears to be important for skin illness classification since informative features can be either small types of texture or large structures of lesions. Nevertheless, having different scales in one network stage does not mean that pieces of information learned at various CNN stages will be useful together. Thus, it is motivating to deal with the cross-scale feature integration.

2.3 Cross-Scale Combination of Features

Extraction of features for multiple spatial scales provides representation with a variety of receptive fields, however, feature extraction from all subsequent layers is not sufficient to prove that information is properly combined. The authors of the original study proposed to use cross-scale feature combination as an additional way of integrating outputs from different layers of the convolution neural network. According to the architecture suggested, first the authors project the outputs of the three layers into one 256-channel space and spatially align them, which results in a concatenated vector of 768 channels that is then reduced to 256 channels with the help of 1×1 convolution. Thus, the lower, mid, and upper feature representation levels make contribution to a unified feature representation at the same time keeping the number of channels reasonable.

Cross-scale fusion is assessed separately in the progressive ablation framework. Here, M1 stands for the stage of extracting features at several scales; M2 brings in the aspect of cross-scale fusion. Through this arrangement, the impact of fusion can be assessed independently from the later addition of CBAM and classifier-head normalization.

2.4 Attention and Explainability

The attention mechanism is a way to allow the convolutional neural network to boost relevant feature channels and spatial zones. The Convolutional Block Attention Module (CBAM) uses channel attention first, then spatial attention, creating a lightweight tool for improvement of learned feature representations [5]. This research employs CBAM as an additional module in the ablation sequence, assessing whether attention-based enhancement brings any improvement compared to multi-scale extraction method and methods of cross-scale combination of features.

Models' interpretability is another crucial component of the classification of medical images because the presence of high prediction accuracy is insufficient to guarantee that the model processes the clinically significant image parts. In this case, Gradient-weighted Class Activation Mapping (Grad-CAM) helps researchers by visually indicating the parts of the images that contributed to making the prediction [6]. Therefore, the authors use Grad-CAM to observe the locations in the image where the CNN models place emphasis and to gain insights about how they make their decisions.

The integration of explainability and attention provides complementary information – CBAM works as a feature-refining tool in the network while Grad-CAM is employed after the predictive process

is completed to obtain information about the areas of interest for the model. These investigations are considered supplementary assessments rather than actual proofs of clinical validity, , a distinction emphasized in recent meta-analytic and systematic-review work comparing AI and clinician performance and examining explainable deep learning for dermatological image diagnosis more broadly [14,15].

2.5 Research Gap

The ability of deep learning, multi-level feature extraction, attention mechanisms, and explainability methods in the analysis of dermatological images has been proven in previous studies. Nevertheless, these elements are mainly studied separately from each other or within complex constructions, thus making it impossible to assess their individual contribution. This study fills this gap with the help of methodological approaches and creates an ablation design in which five CNN models are evaluated. Each of these models involves different steps in terms of multi-scaling feature extraction, cross-scaling feature fusion, CBAM attention application and classifier-head normalization and is therefore assessed separately as one of the methods applied. Other important parameters of classification performance, which are also considered in this study, include model reliability, interpretability, statistical comparison, and dataset integrity. The exact-duplicate check, which is required in this case, is of particular significance as overlapping images have been revealed within the datasets and require extra caution in interpretation of the results.

3. Materials and Methods

3.1 Overall Study Design

In the experiment, controlled design was adopted to study the multi-scale and cross-scale feature acquisition for the classification of skin conditions into five categories with the help of convolutional neural networks. With time, five CNN models were constructed with each of the models incorporating one more component architecture. The process began with a conventional CNN (M0), and then a multi-scale feature extraction was implemented (M1), then the cross-scale feature fusion was developed (M2), then the CBAM attention technique was used (M3), and the last step was the implementation of batch normalization in the classifier head (M4).Thus, the progressive approach gives an opportunity to determine influence from just one change in architecture and still perform the same classification task with several models.

Table 1. Progressive Model Architecture and Main Modifications

Model	Main modification
M0	Baseline CNN
M1	Multi-scale feature extraction
M2	Multi-scale extraction + cross-scale fusion
M3	M2 + CBAM attention
M4	M3 + batch normalization in classifier head

3.2 Dataset and Class Distribution

The dataset comprises of 9,478 photographs that belong to five categories of skin disorders: Nail psoriasis, SJS-TEN, Vitiligo, Acne, and Hyperpigmentation. The dataset is segregated into three sections: training, validation, and testing, which consists of 6,631, 1,421, and 1,426 images respectively, reflecting almost 70%, 15%, and 15% of the total dataset. The training part is utilized for training the model, while the validation section is used for picking the best model, while the testing section is used for measuring the efficacy of the model. Unlike widely used public dermoscopic benchmarks such as HAM10000 [7], the present dataset was assembled specifically for this study and covers five distinct dermatological conditions.

Due to the difference in the distribution of classes among the five categories of skin disorders, class-weighted cross-entropy loss function was used during the training of the neural network. Other evaluation criteria such as balanced accuracy and averaged metrics were also computed to get a better understanding of the performance of the model when there is a difference among the sizes of classes.

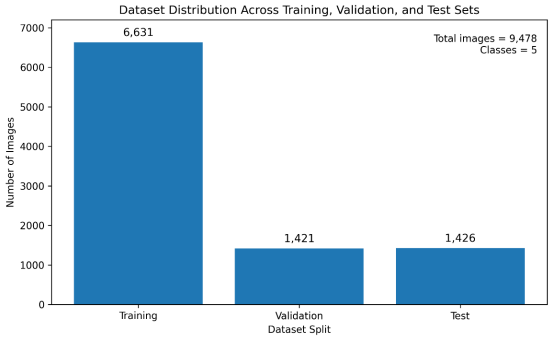

Figure 1. Distribution of images across the training, validation, and test subsets.

There are 9,478 images in the dataset that are classified into five skin disease categories, namely Nail psoriasis, SJS-TEN, Vitiligo, Acne, and Hyperpigmentation. The data is divided into 6,631 training, 1,421 validation, and 1,426 test images. The proportions of data are roughly equal to 70.0%, 15.0%, and 15.0% in the training, validation, and testing stages, respectively. The initial image set was used to train the model, the second set was to choose between models, and the third set was for

assessment of the outcomes. The proportions of the data subsets are shown in Figure 1.

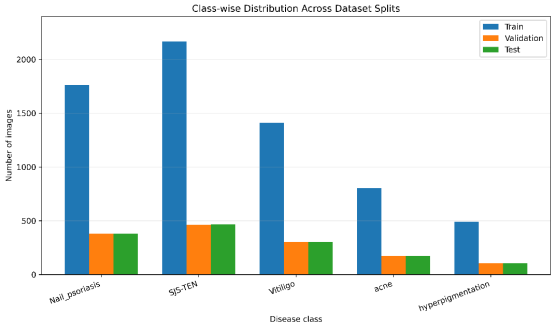

Figure 2. Image counts of skin diseases by class in training, validation and test datasets.

From the figure 2, the ratios of the classes are very much consistent in size but there are significant class size differences at the same time. Hence, it justifies the use of class-weighted loss and balanced accuracy methods in the presented experiments.

3.3 Data Preparation and Augmentation

Before being used in the CNN models, preprocessing steps were implemented for all images through a standardized preprocessing procedure. Data augmentation was utilized for the training set to create more variance in the training sample and decrease the chances of overfitting. The validation set and test samples that were set aside were processed without the use of training-time augmentation to evaluate the model more faithfully. Since the five classes of diseases have not been represented well in the training set, class-weighted cross-entropy loss function was used for optimization purposes. Class weights were found based on the frequencies of classes available in the training dataset, which means that lower weights were assigned to common classes, while rare classes were given higher weights. The resulted distribution of the training set and corresponding class weights is presented in Table 2.

Table 2. Training-set class distribution and corresponding class weights

Class	Training	Class
Nail_psoriasis	1,763	0.7522
SJS-TEN	2,165	0.6126
Vitiligo	1,411	0.9399
Acne	803	1.6516
Hyperpigmentation	489	2.7121

3.4 Model Architecture

The five CNN architectures were created in a sequential manner so that the role of each architecture function could be analysed within the same five-class classification challenge. The basic architecture is provided by model M0. Model M1 develops multi-scale feature extraction. The architecture in M2 is built on that in M1 by including cross-scale feature fusion. Model M3 includes CBAM attention mechanism, while in model M4 batch normalization is implemented in classifier head. The number of parameters ranges from 340,389 in model M0 to 1,338,695 in model M4, which enables researchers to assess accuracy together with improving complexity of the architectures.

3.4.1 Multi-Scale Feature Extraction

Three parallel layers applying different receptive fields, namely 3×3 convolution, 5×5 convolution, and dilated 3×3 convolution, are used in multi-scale feature extraction blocks. After combining the 3 resultant feature maps, they are processed through a 1×1 convolution. There is also an advantage of one additional data input from the input representation through a residual projection. Besides that, batch normalization and SiLU activation functions are used in the feature extraction blocks. With this design, the network can process both local and wider spatial patterns at the same time. During the progressive experiment, this multi-scale element is introduced in M1, and it is the first change of the architecture in comparison with M0.

3.4.2 Cross-Scale Feature Fusion

M2 introduces cross-scale feature fusion, which is the addition of the features produced at various levels within the network. The features are transformed independently from the shallow, middle, and deep levels of the network into a common 256-channel geometry and properly aligned before being concatenated together. As a result, a 768-channel representation is produced and is compressed back to 256 channels afterward via 1×1 convolution. Thus, the final representation has features from shallow, middle, and deep levels combined. It was tested independently from the multi-scale extraction component since the shift from M1 to M2 is the switch to cross-scale fusion.

3.4.3 Attention and Classification Head

The introduction of CBAM in M3 is carried over in M4. The module works by making use of channel attention and subsequently spatial attention, which makes the model more effective in creating a feature representation by allocating more weights to the useful features.M4 is an improved version of M3 as it uses batch normalization in the classification head. Hence, the configuration can be summarized as follows:

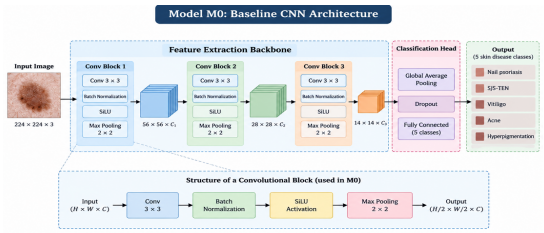

Figure 3: Model M0 Architecture, M0 Baseline CNN Architecture

M0 is the first architecture of CNN. M0 serves as the platform to compare other advanced models. M0 follows traditional sequential constrains by adding layers of convolutions with increasing number of channels to reduce spatial dimension through max pooling. At the end of this procedure, the feature map is transformed by means of average pooling and classification.

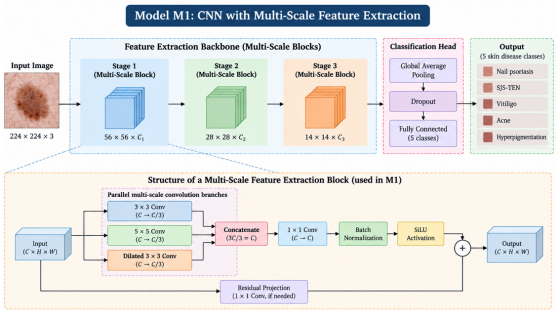

Figure 4: Model M1 Architecture, CNN with Multi-Scale Feature Extraction

M1 is the M0 model with untraditional channels. M1 is the extend of M0 by replacing architecture framework of CNN with multi-scale feature extraction modules. Each of the multi-scale modules has parallel architecture of filters with convolutions layers of different sizes including dilated convolution with dilation rate of 2. The outputs of every branch are connected and pass through another layer of artificial neuron.

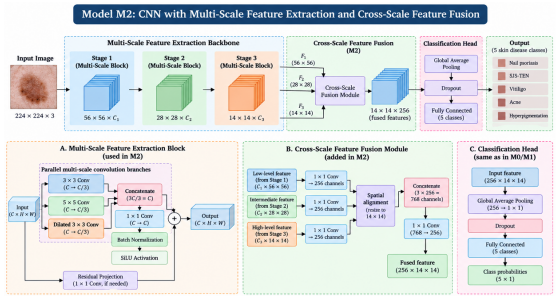

Figure 5: Model M2 Architecture, CNN with Cross-Scale Feature Fusion

The M2 multi-scale CNN (Convolutional Neural Network) model, also known as M2, is an evolution of model M1 (another multi-scale CNN). It features enhancements in its architecture, allowing it to use feature representations from three different levels of the network instead of two. The input feature maps obtained at levels of 64, 128, and 192 channels are transformed separately into 256-channel feature maps using 1×1 convolutions, aligned spatially, and then concatenated to form a 768-layer representation. A 1×1 convolution layer is then employed to reduce the dimensionality of the resulting concatenated feature representation, yielding input for classification processes.

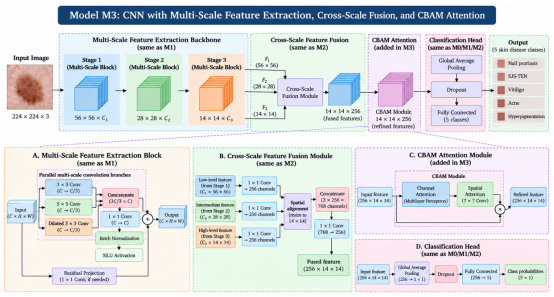

Figure 6: Model M3 Architecture, CNN with CBAM Attention

Another model called M3 (multi-scale CNN based on M2), extends M2 by incorporating CBAM (Convolutional Block Attention Module) after cross-scale feature fusion. The main idea behind CBAM is to sequentially apply channel-wise and spatial attention to the resulting feature representation to further refine the feature before applying global average pooling for making classification decisions. This way, this model enables observation of the additional effects produced by the attention-based refining feature implementation in a progressive architecture.

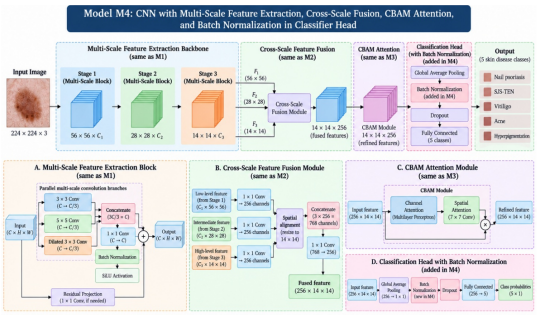

Figure 7: Model M4 Architecture, Convolutional Neural Network (CNN) with Batch Normalization at the Classifier Head

M4 model which is a multi-scale CNN with cross-scale fusion and attention mechanism CBAM consists of batch normalization at the classifier stage. It is worth noting that instead of going directly to classifier stage, additional processing is performed on the feature representation of size 256 by applying global average pooling, batch normalization and other steps to produce the final classification result.

3.5 Training and Optimization

All five CNN architectures were trained from scratch under the same training and validation dataset conditions to conduct a controlled comparison based on architectural variations. The training process for the models was performed using the class-weighted cross-entropy loss and the AdamW optimizer with an initial learning rate of 1×10^{-4} and weight decay of 1×10^{-4}. The learning rate scheduler used here is ReduceLROnPlateau, which is applied based on validation balanced accuracy with a decrease factor of 0.5 and a patience of 2 epochs. Training was done with batch size 32. The best model was selected

based on validation balanced accuracy score and the early stopping rule was applied in case validation accuracy score does not improve for four consecutive epochs. In these experiments, the maximum number of epochs allowed for each model was limited to 25. There was no transfer learning done and the training of models was from scratch.

Table 3. Training and optimization settings

Parameter	Setting
Optimizer	AdamW
Initial learning rate	1×10^{-4}
Weight decay	1×10^{-4}
Batch size	32
Maximum epochs	25
LR scheduler	ReduceLROnPlateau
Reduction factor	0.5
Scheduler patience	2 epochs
Early-stopping	4 epochs
Model-selection	Validation balanced

3.6 Evaluation Metrics and Statistical Analysis

The performance of the models was assessed on the separate test set using accuracy, balanced accuracy, macro precision, macro recall, Macro-F1, sensitivity, and specificity. Additionally, the confusion matrices were analysed to study class-related predictions and common classification errors. Because all five models made predictions for common observations of the test set, McNemar's test allowed for a paired comparison of the predictions. Furthermore, metrics of prediction confidence and Expected Calibration Error (ECE) were analysed to evaluate the relationship between predicted probabilities and actual performance. The methods of Grad-CAM and perturbation-based analysis were utilized as supplementary methods to study the interpretability of the models.

4. Results

4.1 Overall Performance of the Models

The effectiveness of five CNN variants was assessed utilizing a set of 1,426 images held back specifically for testing. Classification outcomes varied considerably among the new architectures.M0, the baseline CNN, has achieved an accuracy of 74.40%, a balanced accuracy of 77.90%, and a Macro-F1 score of 74.19%. M1 has utilized multi-scale feature extraction and achieved an accuracy of 90.60%, a balanced accuracy of 92.45%, and a Macro-F1 score of 89.68%. M2, implementing cross-scale feature fusion, has achieved an increased accuracy of 93.76%, a balanced accuracy of 94.74%, and a Macro-F1 score of 92.96%. M3, which incorporated the CBAM attention, has attained the maximum testing performance of 95.86% accuracy, 96.93% balanced accuracy, and 95.13% Macro-F1 score. M4 has employed batch normalization on the top layer of the classifier and achieved an accuracy of 94.81%, a balanced accuracy of 95.63%, and a Macro-F1 score of 93.94%.

Table 4. Overall test performance of the five CNN models

Model	Accuracy	Balanced	Macro-
M0 Baseline CNN	74.40%	77.90%	74.19%
M1 Multi-Scale	90.60%	92.45%	89.68%
M2 Multi-Scale +	93.76%	94.74%	92.96%
M3 Multi-Scale +	95.86%	96.93%	95.13%
M4 Full Proposed	94.81%	95.63%	93.94%

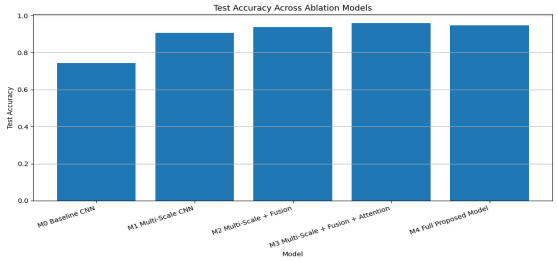

Figure 8. Test Accuracy across Ablation Models

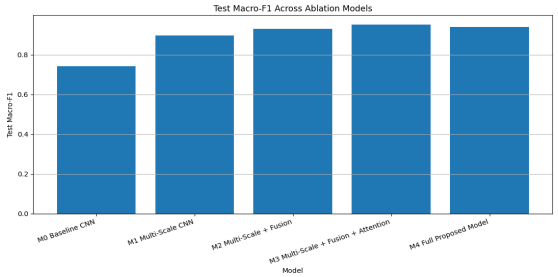

Figure 9. Test Macro F1 across Ablation Models

4.2 Class-Wise Performance and Confusion Matrix Analysis

Class-wise evaluation was done to analyse the behaviours of the best-performing M3 model for the five disease categories. For every class, M3 proved to be quite effective, particularly with respect to the recall achieved, with the highest recall being achieved with respect to acne (100.00%) and hyperpigmentation (100.00%). This was followed in terms of recall by Nail_psoriasis (98.15%), Vitiligo (93.40%), and SJS-TEN (93.12%). As far as the F1-score results are concerned, nail_psoriasis had 97.51%, SJS-TEN had a score of 96.22%, Vitiligo had a score of 94.81%, and hyperpigmentation had 90.21%, while also having the lowest precision of 82.17%. The implications here are that while hyperpigmentation was detected correctly in all cases, some of the predictions were problematic in that images belonging to different categories were

predicted as hyperpigmentation. Overall class-wise results in this study show that while there were only minor variations across all five disease categories, SJS-TEN and hyperpigmentation remain the most difficult cases.

Table 5. Class-Wise Classification Performance of M3 on the Held-Out Test Set

Class	Precisi	Recall	F1-	Suppo
Nail_psoriasis	96.88%	98.15 %	97.51 %	379
SJS-TEN	99.54%	93.12 %	96.22 %	465
Vitiligo	96.26%	93.40 %	94.81 %	303
Acne	94.02%	100.00 %	96.92 %	173
Hyperpigmentat ion	82.17%	100.00 %	90.21 %	106
Macro average	93.77%	96.93 %	95.13 %	1,426

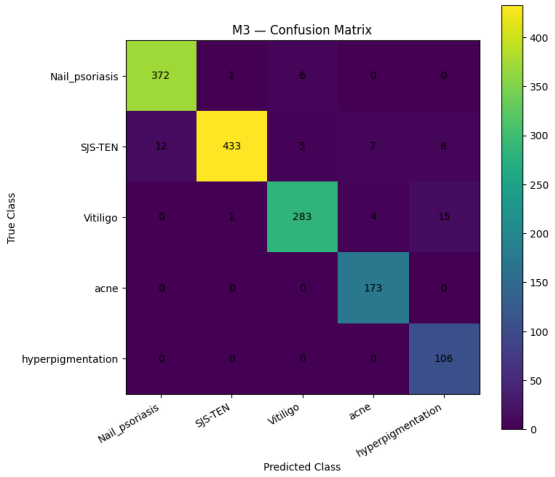

Figure 10. Confusion Matrix of M3

4.3 Statistical Comparison Between Different Predictions from Models

The following chapter will concentrate on the use of the McNemar's test instead of reiterating the performance measures.Given that all the five models are evaluated with the same test data, the McNemar's test has been applied for comparing their predictions. The result obtained for M2 and M3 is significant with p = 0.00352 indicating that the difference in their results is significant at the level of significance of 0.05. As a result of the experiment conducted, M3 has shown higher accuracy, balanced accuracy, and Macro-F1 score as compared to M2. Thus, the statistical comparison brings more proof to confirm that the difference in terms of prediction the two models has displayed is not only related to the overall performance indicators alone.

Table 6. McNemar's test comparison

Model comparison	McNemar's p-value	Significance
M2 vs. M3	**0.00352**	Significant (p < 0.05)

4.4 Analysis of Prediction Confidence

An analysis was performed on prediction confidence for the model M3 to determine if correct and incorrect classifications by the model resulted in the same levels of prediction confidence. The average prediction confidence for correct classifications was equal to 0.9587 while that of incorrect predictions was lower at 0.7174. This shows that the model assigned higher confidence to its correct predictions than to its incorrect ones, but there were still some wrong predictions that occurred with relatively high levels of confidence thereby suggesting that high prediction confidence does not guarantee correct classification. The range of prediction confidence for correct and incorrect predictions is illustrated in the Figure 11.

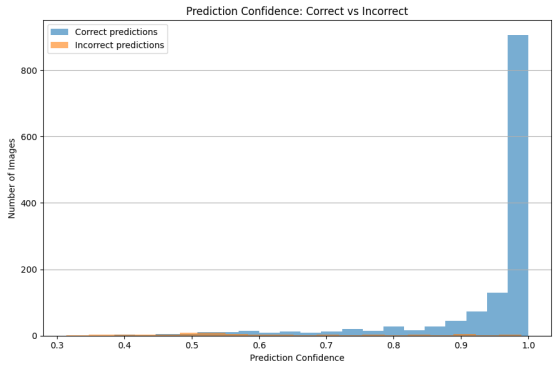

Figure 11. Prediction Confidence Distribution for Correct and Incorrect Predictions of M3

4.5 Model Calibration Analysis

The model calibration was scrutinized to assess the correspondence between the anticipated reliability and actual classification accuracy. The reliability diagrams for M2 and M3, along with the diagonal line representing ideal calibration, are illustrated in Figure 12. The discrepancy between the two models and the perfect calibration line demonstrates that neither model has an ideal level of calibration. Nonetheless, the calibration metric values indicate that M3 has better-calibrated results than M2 in terms of calibration error. For M3, the Expected Calibration Error (ECE) is 0.009898, whereas it is 0.022561 for M2, and the Brier score for M3 is also lower (0.066186 versus 0.096442 for M2). Lower ECE and Brier score signify better agreement between the predicted probabilities and actual outcomes in the experiment.

Table 7. Calibration Performance of M2 and M3

Model	ECE ↓	Brier
M2 Multi-Scale + Fusion	0.022561	0.096442
M3 Multi-Scale + Fusion + Attention	0.009898	0.066186

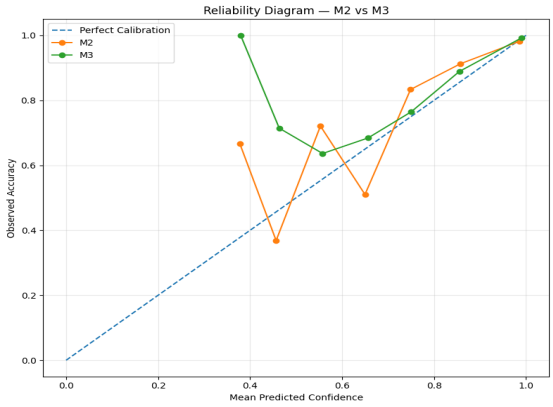

Figure 12. Reliability Diagram Comparing M2 and M3

4.6 Explainability Analysis via Grad-CAM and LIME

To understand which visual areas are responsible for the model's decision making, the Grad-CAM method was deployed in relation to the M3 model. The Grad-CAM analysis utilized the last multi-scale feature block (block3) as its target layer and produced activation maps for sample images from the test set. The examples used in the analysis consisted of those samples which were correctly classified by the model and those samples which were not classified properly.

In the case of a representative correctly classified example, the model's prediction was Nail_psoriasis with a confidence level of 0.9939. Moreover, Grad-CAM was also used for one representative test image in each of the disease classes. The resulting images served as qualitative illustrations of spatial areas related to the model's predictions.

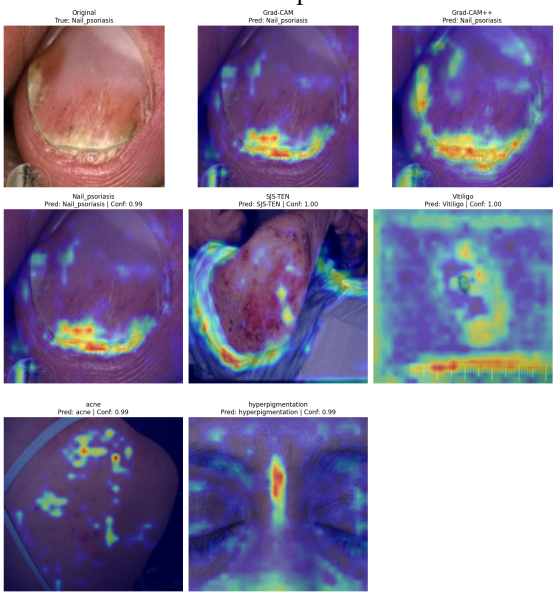

Figure 13. Grad-CAM Visualizations for Representative Test Images Across the Five Skin Disease Classes

LIME Explainability Analysis

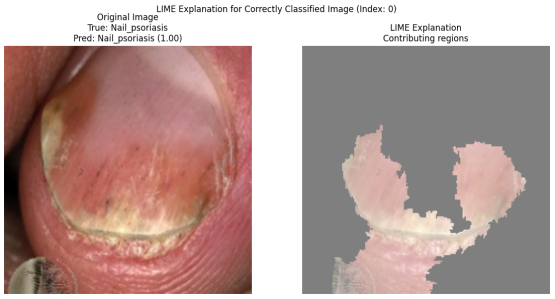

Figure 14. LIME Explanations for the Correctly Classified Image from Nail Psoriasis Category

Along with the Grad-CAM analysis, Local Interpretable Model-Agnostic Explanation (LIME) was used to identify the image areas contributing to the model's predictions. LIME is a local explanation method where the regions of the image leading to a certain prediction for a given input are being known. An illustration of a correctly classified image from the Nail_psoriasis category is shown in figure 13. The M3 model was able to predict the image of Nail_psoriasis with confidence equal to 1.00, while LIME explanation discovered that the decision was made based on regions within and around the nail area. This illustration demonstrates the additional local interpretation of the model's decision and shows how LIME can be used to visualize the areas of the image responsible for specific classification, as seen in figure 14.

4.7 Perturbation-Based Robustness Analysis

To understand how stable the model predictions are, robust measurements were performed. The Grad-CAM methods were checked for reliability with controlled perturbations, for which three cases were chosen: changing brightness, horizontal flipping, and small rotation. Consistency of results obtained with Grad-CAM methods was estimated by mean values of Intersection over Union (IoU) coefficient that is obtained in comparing the original Grad-CAM maps and those obtained under perturbation influence. The prediction confidence was checked before and after perturbations were applied.

Changing brightness managed to produce the most consistent results, as indicated by the general average IoU scores obtained – IoU=0.9220 ± 0.0375, compared to small rotation producing such IoU=0.5312 ± 0.1060, and horizontal flipping producing the least consistent results - IoU=0.2714 ± 0.1580. It is important to note that the prediction certainty was steady throughout the perturbation process. The confidence values were as follows: the original value – 0.9750, the brightness alteration resulted in making it 0.9740, horizontal flipping – 0.9634, and small rotation – 0.9766. Based on these results, it is safe to conclude that the model is using the most reliable efficiency with changing brightness perturbations, while the spatial transformations had their biggest influence on the areas of interest in the images, which included horizontal flipping.

Table 8. Grad-CAM Robustness and Prediction-Confidence Stability Under Input Perturbations

Perturbation	Mean IoU	Std. IoU	Original Confidence	Perturbed Confidence
Brightness change	0.9220	0.0375	0.9750	0.9740
Horizontal flip	0.2714	0.1580	0.9750	0.9634
Small rotation	0.5312	0.1060	0.9750	0.9766

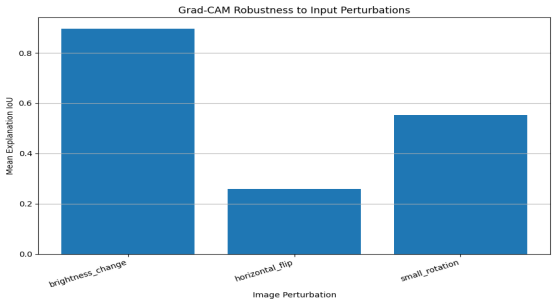

Figure 15. Robustness of Grad-CAM Explanations to Input Perturbations

4.8 Quantitative XAI and Explanation Agreement Analysis

In order to measure the degree of consistency between the model explanations, Grad-CAM and Grad-CAM++ were compared by means of the intersection over union (IoU). As seen in Table 9, the two explanation algorithms revealed different levels of spatial agreement when compared across all five types of disease. The greatest agreement was recorded in the case of SJS-TEN (IoU = 0.7069); a lot less agreement was noted with respect to acne and hyperpigmentation, with IoU values of 0.1442 and 0.1854 accordingly. As shown by the numbers obtained, the explanation patterns can change depending on the disease type and visualization techniques should not be treated as valid indicators of the model's reasoning processes.

Table 9. Agreement Between Grad-CAM and Grad-CAM++ Explanations

Class	Grad-CAM Prediction	Grad-CAM++ Prediction	Grad-CAM Confidence	Grad-CAM++ Confidence	Explanation IoU
Nail_psoriasis	Nail_psoriasis	Nail_psoriasis	0.9939	0.9939	0.4347
SJS-TEN	SJS-TEN	SJS-TEN	0.9988	0.9988	0.7069
Vitiligo	Vitiligo	Vitiligo	0.9964	0.9964	0.4085
Acne	Acne	Acne	0.9950	0.9950	0.1442
Hyperpigmentation	Hyperpigmentation	Hyperpigmentation	0.9853	0.9853	0.1854

5. Discussion

5.1 Performance of the Model

This progressive evaluation showed that applying multi-scale feature extraction resulted in much higher performance than that of baseline CNN. M1 was able to raise its accuracy from the original 74.40% to 90.60%. In contrast, the new model M2, including the cross-scale feature fusion technology improved performance further and its accuracy reached 93.76%. Finally, the introduction of CBAM gave the best overall model performance in the experiment conducted with the model named M3 recording 95.86% classification accuracy, 96.93% balanced accuracy, and 95.13% Macro-F1 scored. The last model M4 did not improve performance anymore and is an indication that having additional classifier-head normalization was not helpful in this specific experiment.

5.2 Performance by Classes and Model Stability

M3 showed great performance with respect to all five classes and has especially high recall for the acne and hyperpigmentation classes. However, hyperpigmentation also showed low precision being equal to 82.17%. Additionally, the model was better in terms of prediction confidence in correct prediction events (0.9587) compared to incorrect prediction events (0.7174). The ECE (0.009898) and Brier score (0.066186) for the model were also lower than in model M2, which is a sign of the probability calibration improvement.

5.3 Insight and Dependability

Grad-CAM and LIME offered complementary visual interpretations regarding the predictions made by the model. The alteration investigation indicated that there is a variation in the stability of

explanations depending on the transformation that the input undergoes. The brightness transformations were the ones that produced the highest average value in terms of IoU (0.9220) and were followed by the angle transformations (0.5312) and the horizontal flipping transformation (0.2714). This means that the model's explanations were found to be stable under brightness transformations, but more fragile during spatial transformations. Therefore, it can be concluded that the combination of classification, calibration, explanation, and dependability analytics has great value in the process of evaluation of a medical image classification model.

6. Conclusion

This research work performed an analysis for a new architecture of CNN for five different types of skin disease classification. The results show a significant improvement of multi-scale feature extraction and feature fusion over the baseline model. It also shows that M3, which itself incorporates CBAM, achieves the best results out of the architectures tested with an accuracy of 95.86%, balanced accuracy of 96.93%, and Macro-F1 score of 95.13% on the test set.

The confidence level of the model, the calibration, and the results of the analyses performed using GradCAM, and LIME were also presented. In conclusion, the study proves the effectiveness of the application of multi-scale feature learning, cross-scale feature fusion, and attention.

7.References

[1] Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. Nature, 542(7639), 115–118. https://doi.org/10.1038/nature21056

[2] Kassem, M. A., Hosny, K. M., Damaševičius, R., & Eltoukhy, M. M. (2021). Machine learning and deep learning methods for skin lesion classification and diagnosis: A systematic review. Diagnostics, 11(8), 1390. https://doi.org/10.3390/diagnostics11081390

[3] Wu, Y., Chen, B., Zeng, A., Pan, D., Wang, R., & Zhao, S. (2022). Skin cancer classification with deep learning: A systematic review. Frontiers in Oncology, 12, 893972. https://doi.org/10.3389/fonc.2022.893972

[4] Qian, S., Ren, K., Zhang, W., & Ning, H. (2022). Skin lesion classification using CNNs with grouping of multi-scale attention and class-specific loss weighting. Computer Methods and Programs in Biomedicine, 226, 107166. https://doi.org/10.1016/j.cmpb.2022.107166

[5] Woo, S., Park, J., Lee, J.-Y., & Kweon, I. S. (2018). CBAM: Convolutional Block Attention Module. In Proceedings of the European Conference on Computer Vision (ECCV) (pp. 3–19). Springer.

[6] Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization. In Proceedings of the IEEE International Conference on Computer Vision (ICCV) (pp. 618–626).

[7] Tschandl, P., Rosendahl, C., & Kittler, H. (2018). The HAM10000 dataset: A large collection of multi-source dermatoscopic images of common pigmented skin lesions. Scientific Data, 5, 180161. https://doi.org/10.1038/sdata.2018.161

[8] Choy, S. P., Kim, B. J., Paolino, A., Tan, W. R., Lim, S. M. L., Seo, J., Tan, S. P., Francis, L., Tsakok, T., Simpson, M., Barker, J. N. W. N., Lynch, M. D., Corbett, M. S., Smith, C. H., & Mahil, S. K. (2023). Systematic review of deep learning image analyses for the diagnosis and monitoring of skin disease. npj Digital Medicine, 6, 180. https://doi.org/10.1038/s41746-023-00914-8

[9] Debelee, T. G. (2023). Skin lesion classification and detection using machine learning techniques: A systematic review. Diagnostics, 13(19), 3147. https://doi.org/10.3390/diagnostics13193147

[10] Hasan, M. K., Ahamad, M. A., Yap, C. H., & Yang, G. (2023). A survey, review, and future trends of skin lesion segmentation and classification. Computers in Biology and Medicine, 155, 106624. https://doi.org/10.1016/j.compbiomed.2023.106624

[11] Lyakhova, U. A., & Lyakhov, P. A. (2024). Systematic review of approaches to detection and classification of skin cancer using artificial intelligence: Development and prospects. Computers in Biology and Medicine, 178, 108742. https://doi.org/10.1016/j.compbiomed.2024.108742

[12] Furriel, B. C. R. S., Oliveira, B. D., Prôa, R., Paiva, J. Q., Loureiro, R. M., Calixto, W. P., Reis, M. R. C., & Giavina-Bianchi, M. (2024). Artificial intelligence for skin cancer detection and classification for clinical environment: A systematic review. Frontiers in Medicine, 10, 1305954. https://doi.org/10.3389/fmed.2023.1305954

[13] Baig, R., Bibi, M., Hamid, A., Kausar, S., & Khalid, S. (2020). Deep learning approaches towards skin lesion segmentation and classification from dermoscopic images: A review. Current Medical Imaging, 16(5), 513–533. https://doi.org/10.2174/1573405615666190129120449

[14] Salinas, M. P., Sepúlveda, J., Hidalgo, L., Peirano, D., Morel, M., Uribe, P., Rotemberg, V., Briones, J., Mery, D., & Navarrete-Dechent, C. (2024). A systematic review and meta-analysis of artificial intelligence versus clinicians for skin cancer diagnosis. npj Digital Medicine, 7, 125. https://doi.org/10.1038/s41746-024-01103-x

[15] Dakhli, R., & Barhoumi, W. (2026). Explainable deep learning for imaging-based skin lesion diagnosis: A systematic literature review. Concurrency and Computation: Practice and Experience.

[16] Surwadkar, T. and Deshmukh, S., 2025. Dimensionality reduction and classification of dermatological images using PCA and machine

learning. *EPJ Web of Conferences*, 341, p.01037. https://doi.org/10.1051/epjconf/202534101037.

[17] Surwadkar, T. and Deshmukh, S., 2026. Interpretable AI for skin cancer prediction using clinical indicators for Indian patients. In: *2026 International Conference on Communication, Computing and Emerging Technologies, (IC3ET)*, pp. 1–9. IEEE. https://doi.org/10.1109/IC3ET64989.2026.11467357

[18] Surwadkar T, Deshmukh S. Interpretable AI for Skin Disease Diagnosis: A Review on Enhancing Trust and Decision-Making in Dermatology. J Neonatal Surgery. 2025 May 12 [cited 2026 Jun. 11]; 14(7): 382-8. Available from: https://www.jneonatalsurg.com/index.php/jns/article/view/5633