# A Hybrid Deep Learning Model for Multiclass Skin Cancer Classification Using ConvNeXtV2 and Separable Self-Attention Mechanisms

مجلة
ميدأوشن
للبحوث والدراسات

ISSN: 3078-5669

Mishaal Mashaan Al-Otaibi

Artificial Intelligence Collage of Informatics,
Midocean University, Comoros.

Received: 06/10/2025
Reviewed: 23/12/2025
Issued at: 15/1/2026

## Abstract

**Objectives:** Skin cancer is among the most prevalent and life-threatening cancers worldwide, making accurate and automated diagnostic approaches essential. This study aims to develop an efficient deep learning–based framework for multiclass skin lesion classification to enhance early detection and support clinical decision-making.

**Methodology:** The study proposes a hybrid deep learning architecture that integrates ConvNeXtV2 convolutional blocks with separable self-attention mechanisms, enabling simultaneous learning of fine-grained local features and long-range contextual dependencies. Transfer learning, extensive data augmentation, and class balancing techniques were employed to improve model generalization and robustness. The proposed model was trained and evaluated on the HAM10000 dataset using five-fold stratified cross-validation to ensure reliable performance assessment.

**Results:** Experimental results demonstrate that the proposed framework outperforms conventional convolutional neural networks and Vision Transformer-based models. The model achieved an average accuracy of 93.52%, precision of 93.17%, recall of 91.24%, and an F1-score of 92.18%, along with a ROC-AUC value of 0.957, indicating strong discriminative capability across multiple skin lesion classes.

**Conclusion:** The findings confirm that combining ConvNeXtV2 with effective attention mechanisms provides a computationally efficient and highly accurate solution for automated skin cancer detection. Future work will focus on multimodal data integration, edge deployment, and clinical validation to enhance real-world applicability and translational impact in healthcare settings.

**Keywords:** Skin Cancer, Deep Learning, ConvNeXtV2, Separable Attention, Transfer Learning, Medical Image Analysis.

## Introduction:

Skin cancer has been reported to be one of the most rapidly increasing malignancies in the world and has still remained a significant burden to healthcare systems. The world health organization (WHO) reveals that every year millions of new cases of skin cancer are diagnosed but more cases of deaths are caused by melanoma although it is less common than non-melanoma skin cancer. The exit stage of the disease is highly important in improving the survival rate, but visual observation of dermoscopic images is still very difficult even among well-trained dermatologists because of minor changes in the texture, color, and other irregularities of the lesions, which may cause misdiagnosis and inter-observer errors (Ozdemir& Pacal, 2024)

Conventional methods of diagnostics based on the dermoscopic examination and then on biopsy and histopathological analysis are lengthy, invasive, and largely reliant on clinical experience. Such constraints are multiplied in resource-starved environments where not many trained dermatologists are available, which underscores the necessity of effective computer-aided diagnostic (CAD) systems. The recent developments in the field of artificial intelligence and especially deep learning-based medical image analysis have shown a lot of potential in both increasing the diagnostic accuracy and decrease clinical workload.

The convolutional neural networks (CNNs) and the Vision Transformer (ViT) architecture have been significantly successful in dermatological image classification by training hierarchical feature representations directly on raw pixel data. However, the traditional CNNs are local receptive field by design, and thus unable to capture long-range contextual dependencies that are important to differentiate visually similar skin lesions. On the other hand, transformer-based models use self-attention mechanisms to learn the global dependencies but at a high computational and memory cost, which restricts their usage in clinical practice.

Recent research has seen the discussion of hybrid architectures as a means to deal with these issues. As an example, (Ince et al., 2025) proposed a ConvNeXtV2-based U-Net system on cerebral vascular occlusion segmentation, showing that ConvNeXtV2 blocks are much more effective in feature extraction with negligible clinical impact. On the same note (Ozdemir& Pacal, 2024), confirmed that ConvNeXtV2 with separable self-attention mechanisms is significantly more effective than CNN-based and ViT-based baselines at multiclass skin lesion classification. These results suggest that the convolutional backbone of ConvNeXtV2 is highly competitive in terms of fine-grained image features of medical images and attention mechanisms can help focus on important diagnostically significant areas.

Although these progressions have been achieved, the current methods usually have high computational cost, most of them do not generalize on unbalanced medical data or lack adequacy in incorporating effective attention mechanisms that are feasible in real-world clinical application. Thus, the lack of a research gap exists in the development of a hybrid deep learning architecture that maintains a tradeoff between classification, computational, and robustness to multiclass skin lesion. To fill this gap, based on the present study, a ConvNeXtV2-based hybrid network with separable self-attention is proposed to obtain reliable, scalable, and clinically feasible skin cancer classification using dermoscopic image data.

## Research Problem:

The main issue of this research consists in correctly and effectively classifying skin lesions with deep learning techniques that can trade-off diagnoses with computational costs. The problem of skin cancer diagnosis is twofold: not only is the morphological heterogeneity of the types of lesions, but also there is a lack of annotated medical images. Dermatologists in clinical practice use the visual examination with the assistance of the dermoscopy, but despite the experience, the boundary between malignant and benign lesions like melanoma, basal cell carcinoma and benign keratosis may be unclear. The nuanced changes in the textual features, the colour differences, and the irregular edges usually cause inter-observer variability, misclassification and late diagnosis.

This diagnostic ambiguity has driven the pursuit of automated systems capable of providing consistent high precision classification in accordance with clinical standards.

Conventional computer-aided diagnosis (CAD) systems are based on manual features that are defined by humans, including texture descriptors, shape indices and color histograms, which are necessarily restricted by human-based heuristics. These methods do not scale to a variety of imaging modalities and cannot realize the complex hierarchical representations that are required to achieve high lesion classification. Various of these limitations have been alleviated by deep learning and CNN-based models, which facilitate the learning of features in datasets with large scale. Nevertheless, CNN architectures are still characterized by major weaknesses. They are constrained by their local receptive fields in their ability to represent longer-range contextual information and are vulnerable to noise, artifacting of illumination, and data imbalance. Moreover, CNNs usually have a tendency to overfit when they use small medical datasets, because of their large number of parameters, and their dependence on large labeled datasets (Farea et al., 2024)

Vision Transformer architectures have also become a promising alternative because it can capture global dependencies by self-attention mechanisms. However, the models generally need huge amount of data and computing power to work optimally, which makes them less usable when dealing with small medical datasets. Moreover, they are too high-dimensional and their attention calculations are too complicated to implement in practice and use in clinical practice. There is therefore an acute need to have hybrid deep learning designs that would allow combining the local feature-elevating abilities of CNNs with the global contextual interpretation of self-attention networks, it would be necessary to ensure computational efficiency.

The research problem in this thesis thus states as follows:

**What is the way to design a hybrid deep learning architecture to obtain an accurate, generalizable and computationally sparse multiclassification of skin lesions with dermoscopic images and combine transfer learning and attention mechanisms?**

## Importance of Research:

### Creation of a Hybrid Deep Learning Framework:

The thesis gives a single model that combines ConvNeXtV2 convolutional blocks during the initial stages of extracting fine-grained local information and applies separable self-attention at the subsequent stages to acquire the global dependencies. This synergy in architecture has been successfully used to close the representational divide between CNNs and transformer models, whilst providing high accuracy without large computational expenses.

### Optimization of Transfer Learning Strategy:

The study uses transfer learning on large-scale natural pictures to dermoscopic pictures, which makes convergence of the study effective despite the small number of samples. The methodology involves upper layers selective fine-tuning and layer freezing of lower layers to preserve generic low-level features which produce better generalization and lower overfitting.

### General Preprocessing and Data Balancing Pipeline:

An effective preprocessing scheme was applied to eliminate the artifact like hair and lighting irregularities with morphological operations and inpainting techniques. Oversampling and augmentation were used to balance the dataset, which removed the problem of imbalance in the classes of clinical datasets, providing fair model training.

### Strict Experimental Design:

The research uses stratified five-fold cross-validation to guarantee an equal estimation of performance and avoids overfitting. Several measures such accuracy, precision, recall, F1-score, and ROC are presented providing a multidimensional perspective of the diagnostic power of the model.

### Empirical Testing and Benchmarking:

Comparative experiments show the proposed hybrid model to achieve a better classification performance on the ISIC 2019 dataset, on average, reaching 93.5% accuracy, and higher F1-scores, and be at the same time computationally efficient than over ten existing CNN and ViT-based architectures. This makes the model one of the most successful frameworks in the field.

## Study Objectives:

- To perform an in-depth study of the current techniques of deep learning in detecting and classifying skin cancer.
- To introduce with a hybrid architecture that combines ConvNeXtV2 blocks, and separable self-attention components.
- To use transfer learning and data augmentation methods to overcome the problem of the lack of annotated data and improve the model generalization ability.
- To compare the proposed framework with the state-of-the-art CNN and ViT models in the case of consistent experimental.

## Theoretical Literature of the Study:

Skin cancer is still one of the most common types of cancers as well as one of the fatal types of cancer across the globe. Diagnosis of skin cancer is a difficult task and this is because of the visual similarities between benign and malignant lesions that result in early diagnosis and hence, the improvement of survival. The dermatologists often use the dermoscopic images to help them diagnose the skin conditions, but since there are no trained medical practitioners to help them, there is increasing need to have automated systems that can help them diagnosis the skin cancer accurately and efficiently. The recent developments in deep learning and transfer learning, especially in medical image analysis, promise to be a significant solution in this problem.

Convolutional neural networks (CNNs) and deep learning methods have already been used to classify and identify skin cancer using dermoscopic images. One technique that has been important in enhancing the performance of deep learning models is transfer learning that enables the utilization of pre-trained weights of large datasets when such are not available in the medical imaging domain due to the low number of labeled data. Application of transfer learning models such as Xception have also been effective in the detection of skin cancer as they allow features to be extracted in the dermoscopic images to enhance the accuracy of classification (Alotaibi & AlSaeed, 2025) .

Nevertheless, a limitation to deep learning models in medical image classification is the capability to concentrate on the most pertinent attributes of the images, including the finer details of the images in terms of their texture, colour, and edge shapes that distinguish benign and malignant lesions. This is where attention mechanism (AMs) has been incorporated into deep learning models to improve its performance. Attention mechanisms enable the model to pay attention to the most significant aspects of the image that enhances classification accuracy, especially in complicated medical imaging tasks such as skin cancer detection.

As an illustration, self-attention enables the model to acquire the connection between various regions of the image, whereas soft-attention randomly distributes attention throughout the image, and hard-attention chooses the areas that are the most pertinent (Ravi, 2022).

Besides, the addition of attention mechanisms by the Xception transfer learning model to the model increased its accuracy in identifying benign and malignant lesions by 94.11% (self-attention), 92.97% (hard attention), and 93.29% (soft attention). Attention mechanisms added as well increased recall metrics especially in medical applications that are life threatening on false negatives. These results highlight the role of attention processes in medical imaging, especially in making the model concentrate on the most important features that are essential in accurate diagnosis (Alotaibi & AlSaeed, 2025).

Other studies have also proposed hybrid models that use CNNs with other machine learning algorithms such as long short-term memory (LSTM) networks, which learn temporal patterns of sequential picture information. The hybrid model like the resnet50-lblstm model combines the advantages of CNNs and LSTMs to process both spatial and sequential information on the images to enhance skill of the model in classifying skin lesions. With the addition of transfer learning, these hybrid models can be specialized to particular tasks, and they are much better at classifying different types of skin cancer with impressive accuracy levels of more than 99% (Mavaddati, 2025).

Moreover, methods to solve the problem of dataset imbalance, which is also a frequent issue in medical image classification, have also been considered. With attention and cost-sensitive learning, the model is able to tackle imbalanced datasets by prioritizing the infrequent classes which is essential to identifying less prevalent and more harmful skin cancers. Multi-model ensemble methods have been demonstrated to enhance the accuracy of classification due to the reduction of the bias caused by the imbalance of data and the increase of the reliability of the predictions (Ravi, 2022)

To sum up, the combination of the attention mechanisms and transfer learning and deep learning models has a significant potential in improving the skin cancer detection. Such methods enhance the priority regarding key characteristics, boost precision, and make the models stronger in clinical environments where timely and correct diagnosis is crucial in enhancing patient outcomes. The current studies in this field hold the hope of coming up with even better and practical aids to dermatologists especially when the resources are limited by utilizing the power of AI in analyzing medical images.

## Review of Previous Studies:

In recent years, there is a plethora of studies on deep learning-based methods of automated analysis of skin lesions to enhance diagnostic accuracy, robustness, and computational efficiency in a variety of clinical environments.

In research (Ozdemir& Pacal, 2024) a combination of taxonomies was created and tested against the ISIC 2019 data consisting of eight different categories of skin lesions. It used data augmentation and transfer learning methods in order to boost robustness and generalization. The experimental findings showed a 93.48, 93.24, 90.70, and 91.82 accuracy, precision, recall, and F1-score respectively and was able to outperform over a dozen CNN-based and Vision Transformer-based architectures under the same experimental conditions. It is worth noting that the model had kept relatively few parameters, making it to be deployed effectively in costly clinical settings in real-time.

In research (Abohashish et al., 2025), a hybrid architecture, which combined time-distributed Convolutional neural network (CNN) layers with Long Short-Term Memory (LSTM) units, was proposed to classify melanoma and non-melanoma with the HAM10000 dataset. The LSTM component was a sequential relationship learned between patches of an image, both spaces and time, and the CNN layers learned discriminative texture and edge features. The trained and integrated framework, which employed regularization methods and learning rate scheduling, was found to be more accurate, more precise, more recalls better F1-score, and better ROC values than CNN-only baselines.

In a bid to solve the problem of multiclass skin cancer diagnosis, (Tahir et al., 2023) introduced the DSCC_Net, which is a CNN-based deep learning model that was tested on the ISIC 2020, HAM100000, and DermIS datasets. SMOTE Tomek sampling was used to reduce the effect of class imbalance. The model was highly performing with an AUC of 99.43, 94.17 accuracy and recall, and 94.28 precision and F1-score of 93.93. DSCC_Net exhibited much greater performance than a number of popular CNN models (ResNet152, VGG16, VGG19, InceptionV3, EfficientNetB0, and MobileNet), demonstrating its relevance to clinical diagnostic support systems.

A complete pipeline based on skin lesion diagnostic, which incorporates preprocessing, segmentation, feature optimization, and classification was proposed in (Khan et al., 2021) . The LCcHIV method was used to improve the visual contrast and then a 10-layer CNN based saliency-based segmentation was implemented. Transfer learning was used to extract discriminative features and an Improved Moth-Flame Optimization (IMFO) algorithm was used to select the best features. Multi set maximum correlation analysis (MMCA) was employed to perform feature fusion and Kernel extreme learning machine (KELM) was employed to perform classification. The framework was found to be 98.70% segmentation accurate on ISBI 2016/2017 and ISIC 2018, and 90.67% classification accurate on HAM10000, indicating good automation and high diagnostic quality.

Study (Alwakid et al., 2022) aimed at increasing the computational efficiency and classification accuracy by using a hybrid pipeline involving deep feature extraction and evolutionary optimization. The methodology involved enhancement of the images, feature extraction with the transfer learning and optimal feature selection with the help of the hybrid Whale Optimization algorithm directed by entropy and mutual information. An adapted Canonical Correlation Analysis was used in achieving feature fusion, and an Extreme Learning Machine (ELM) used in classification. HAM10000 and ISIC 2018 results provided accuracies of 93.40% and 94.36, respectively, and compared to other existing state-of-the-art methods, it shows better performance.

A deep ensemble methodology was suggested in (Thwin& Park, 2024) to differentiate between basal cell carcinoma, squamous cell carcinoma, and melanoma. The ensemble used three prepared CNNs, namely VGG16, InceptionV3, and ResNet50 whose output was averaged with a weight to utilize the complementary representations of the features. Class imbalance was dealt with using the oversampling methods. The ensemble had an accuracy of 91 percent on unbalanced data and 97 percent on balanced ISIC 2018 data, and overall, similar patterns can be studied on the HAM10000 dataset, indicating high efficiency of ensemble learning and data balancing.

To enhance the accuracy and interpretability, (Efat et al., 2024) also introduced a multi-level deep learning ensemble which involves personalized transfer learning and triple attention attention, such as channel attention, squeeze-and-excitation attention, and soft attention. The framework which was trained on the HAM10000 dataset presented a Multi-Level Information Gain Proportioned Averaging (ML-IGPA) strategy to combine predictions in an optimal way. The model scored at 94.93, which is better than the available methods. Also, Grad-CAM visualizations were used to point out the diagnostically significant areas of lesions, which contributes to transparency and clinical trust in the model.

A hybrid deep learning system was introduced in (Gomathi& Arunachalam, 2024) combining enhanced preprocessing, segmentation, feature mining, and optimized classification. Pre-processing was done by hair removal and median filtering, then U-Net was used to segment lesions. RGB histograms and Gray-Level Co-occurrence Matrix (GLCM) were used to extract color and texture features. The classification was done with a modified LSTM (MLSTM) whereby the architecture was optimized using a Sewing Training-Based Optimization (SC-STBO) algorithm. The model when tested on ISIC data sets reported accuracy and recall of 99.31% and 98.25 respectively when tested on seven types of lesions and performed very well when compared to the traditional methods.

In (Manzoor et al., 2025) , the authors have suggested a dual-stage deep learning pipeline that can be used to annotate the skin lesion diagnosis process in a sequence-based manner by combining segmentation and classification steps. The initial phase used a U-Net architecture to effectively outline the boundaries of lesions using ISIC 2018 dataset. The second stage employed EfficientFormerV2 and SwiftFormer, which are lightweight models built on transformers, to train on HAM10000 dataset to classify lesions. To increase the strength, the models were tested on balanced as well as imbalanced data by performing extensive data augmentation. EfficientFormerV2 was even more efficient with a F1-score of 97.14, a sensitivity of 96.85 and a specificity of 96.70 on the balanced HAM10000 dataset. The framework also illustrated good segmentations with an accuracy of 97.59, Jaccard index of 89.12, and Dice coefficient of 94.24% on ISIC 2018. The model also demonstrated good generalization in the case of teledermatology-like processes, which is demonstrated in its performance in the ISIC 2024 SLICE 3D challenge.

The paper Halder et al., (2025) proposed a fuzzy rank-based ensemble mechanism to improve the strength of skin lesion. Three pretrained CNN models were included in the proposed framework, which consisted of Xception, InceptionResNetV2, and MobileNetV2, and the fusion of the predicted models was carried out through a fuzzy ranking system, which was depending on the confidence of a model and its relative success. The HAM10000 dataset was used to solve the problem of class imbalance through data augmentation. The experimental findings proved that the ensemble was always more active than its components and it could achieve high accuracy, sensitivity, and specificity and computational efficiency, thus it can be applied in automated dermatological diagnostic systems.

A model of lightweight deep learning was introduced in (Wang et al., 2023) with the use of which smart healthcare applications are to be implemented, paying particular attention to the efficiency of calculations and the ability to run on resource-limited processors. The model applied entropy-based weighting and first-order cumulative moment (EW-FCM) lesion-background separation and then used a modified wide-ShuffleNet architecture to perform classification. The number of parameters necessary to fit the proposed model was much smaller (about 79 times less than VGG19) and gave similar or better accuracy. The assessments on the HAM10000 and the ISIC 2019 datasets proved its suitability in mobile and embedded health care settings.

Study Ahmad et al., (2023) introduced a hybrid diagnosis model that combined the image super-resolution preprocessing and deep pretrained CNN models. DenseNet201 was used as the main feature extractor and to extract the fine-grained texture details, Histogram of Oriented Gradients (HoG) descriptors were added. The stage of super-resolution preprocessing improved the quality of low-quality dermoscopic images before classification, which resulted in increased feature discriminability. Benchmarking experiments conducted on experimental data revealed that the designed solution was significantly more successful and effective than the traditional CNN-based classifiers without super-resolution in terms of accuracy and robustness.

A hybrid classification architecture that combines two pretrained CNN models, Xception and ShuffleNet, was developed

in Alzakari et al., (2024) to obtain complementary deep features of dermoscopic images. Before feature concatenation, global average pooling has been used, and then based on an Improved Butterfly Optimization Algorithm (IBOA) optimal feature selection has taken place. Lesion prediction was then done by a lightweight classifier. Grad-CAM visualization was also included in the framework to enhance the interpretability of the framework by identifying diagnostically relevant areas. The accuracy of the ISIC 2018 and HAM10000 datasets was found to be 99.3% and 91.5% respectively, and with a lower computational complexity.

In Hoang et al., (2022) the multimodal deep learning model MiSC was proposed which combined the dermoscopic image features with the metadata of the patients in order to generalize the classification of six types of skin cancer. MobileNetV2 was used to extract image-based features, whereas a Random Forest classifier was used to analyze the clinical metadata. Last predictions were produced due to the hybrid fusion of image and tabular features. The framework tested on the PAD-UFES-20 dataset had an accuracy, 95.6%, a precision, 96.8% and a recall, 95.6%, and a F1-score, 95.7% that exceeded that of image-only models and underlines the advantages of multimodal integration in dermatological diagnosis.

A self-supervised knowledge distillation system, SSD-KD, was proposed in Mohamed et al., (2025) to train a small but high-quality skin lesion classifier. The composite teacher network that includes ResNet152V2, ConvNeXtBase and ViT-Base models in it transferred knowledge regarding instances and between-instances within the network to a small student network. This method has allowed the student network to learn rich feature representations that have low computational complexity. The distilled model is able to reach about 85 percent accuracy on the ISIC 2019 dataset indicating that it can be deployed in resource-restricted clinical environments with little performance loss.

## Summary table

**Table (1): Previous Studies summary**

| Study / Model | Dataset(s) | Main reported accuracy | Key notes |
|---|---|---|---|
| ConvNeXtV2 + separable attention (hybrid) | ISIC 2019 | 93.48% | Hybrid model; small parameter count; outperformed >10 CNN & >10 ViT baselines. |
| Time-distributed CNN + LSTM | HAM10000 | (reported ↑ vs CNN baselines) | Patch-sequence modeling with LSTM; improved ROC/F1 |

| | | | |
|---|---|---|---|
| | | | over CNNs (exact %s in original). |
| DSCC_Net (CNN) | ISIC 2020, HAM10000, DermIS | 94.17% (ISIC 2020) | SMOTE-Tomek for balancing; AUC 99.43%; outperformed ResNet152, VGG, EfficientNetB0. |
| LCcHIV + 10-layer CNN + KELM pipeline | ISBI 2016/2017, ISIC 2018, HAM10000 | 90.67% (HAM10000 classification) / Segmentation 98.70% | Fully automated pipeline; IMFO feature selection; strong segmentation results. |
| Deep features + Whale Optimization + ELM | HAM10000, ISIC 2018 | 93.40% (HAM10000) / 94.36% (ISIC 2018) | Evolutionary optimization + feature fusion; competitive state-of-the-art. |
| Ensemble (VGG16, Inception-V3, ResNet50) | ISIC 2018, HAM10000 | 97% (balanced ISIC 2018) / 96% (balanced HAM10000) | Weighted averaging ensemble; large accuracy gain on balanced sets. |
| Multi-level ensemble + triple attention (ML-IGPA) | HAM10000 | 94.93% | Triple attention (channel, SE, soft); Grad-CAM interpretability. |
| U-Net + MLSTM (SC-STBO optimization) | ISIC (various) | 99.31% | Advanced preprocessing + optimized MLSTM; very high reported accuracy. |
| Dual-stage: U-Net segmentation + EfficientFormerV2 classification | ISIC 2018, HAM10000 | 97.59% (segmentation acc.) / 97.11% F1 (balanced HAM10000) | Strong segmentation + lightweight transformer classifiers; good telederm generalization. |
| Fuzzy rank-based ensemble (Xception, InceptionResNetV2, MobileNetV2) | HAM10000 | (higher than individual models) | Fuzzy fusion based on confidence/rank; robust and efficient ensemble. |
| EW-FCM + wide-ShuffleNet (lightweight) | HAM10000, ISIC 2019 | Comparable to heavy nets (≈ high 80s–90s) | Very low parameter count (~79× smaller than VGG19); mobile/embedded focus. |
| Super-resolution + DenseNet201 + HoG | Benchmark datasets | (improved vs non-SR baselines) | SR preprocessing improved feature quality and robustness. |
| Xception + ShuffleNet + IBOA feature selection | ISIC 2018, HAM10000 | 99.3% (ISIC 2018) / 91.5% (HAM10000) | Complementary features + Grad-CAM; lightweight final classifier. |
| MiSC (image + metadata fusion) | PAD-UFES-20 | 95.6% | Multimodal fusion (MobileNetV2 + Random Forest on metadata) → clear gain over image-only. |
| SSD-KD (self-supervised distillation) | ISIC 2019 | ≈85% | Knowledge distillation to create lightweight student model; good trade-off for resource-limited settings. |

## Research Gap and Contribution of the Present Study

Although the current deep learning-based skin lesion classification methods attained significant progress, a number of drawbacks are present. Numerous state-of-the-art approaches are based on computationally costly ensemble architectures, multimodal inputs or complicated attention mechanisms, which cannot be easily deployed in clinical settings in real-time. Lightweight models are efficient, but can also compromise classification accuracy or find it hard to model global contextual dependencies. Moreover, some methods are mostly based on either convolutional feature extraction or transformer-based attention without combining both of the paradigms in a computationally efficient way.

To overcome these shortcomings, the current paper suggests a hybrid deep learning architecture, which will be based on the combination of ConvNeXtV2 convolutional blocks and separable self-attention components. This architecture allows efficient acquisition of small-scale local features and efficient modeling of long-range dependencies, and has a good trade-off between accuracy, understandability, and computational efficiency. The proposed model unlike heavy ensemble or multimodal frameworks has a smaller architecture and still shows better results on the HAM10000 dataset, which makes a step towards scalable and clinically useful automated skin cancer diagnosis models

## . Methodology:

### Overview of the Proposed Method

The main aim of this paper was to come up with a concise and sound deep learning framework to classify skin lesions in multiple classes. The suggested methodology was based on the current concepts of deep learning and transfer learning to solve the issues of medical imaging datasets, such as variability of the visual representations of lesions, imbalance of classes, and imaging artefacts, such as hair covering the face and uneven illumination**.** The generalized workflow involved data acquisition, image preprocessing, data augmentation and class balancing, feature extraction by the use of pretrained backbones, model training, stratified cross-validation, and performance evaluation. The stages were all intended to be highly diagnostic, robust and clinically relevant.

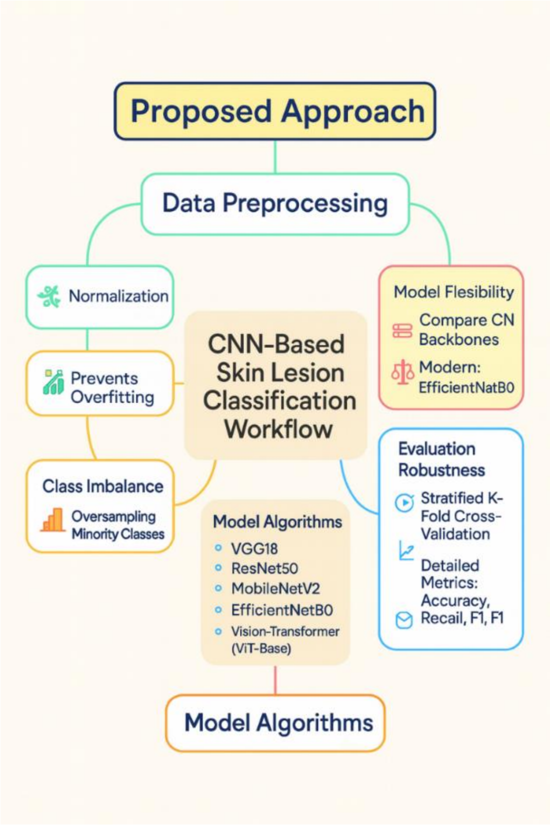

**Figure (1): Proposed CNN-based skin lesion classification workflow illustrating data preprocessing, model architectures, and evaluation strategy.**

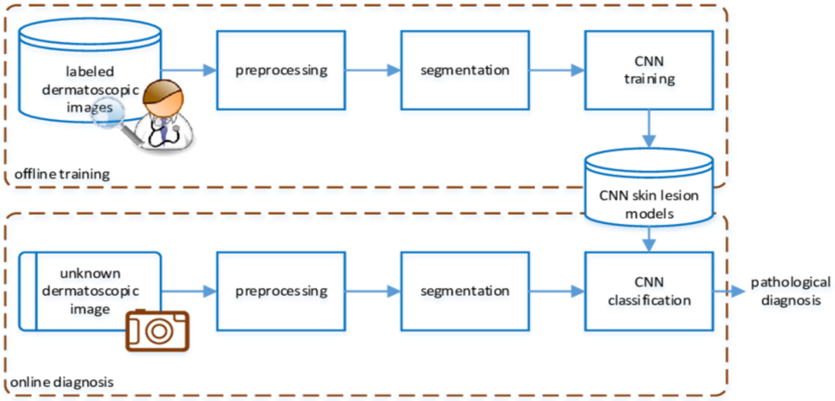

**Figure (2): Proposed Approach**

## Description of the Data and Permissions of Use.

The Skin Cancer MNIST: HAM10000 dataset was used as the experimenter of the Skin Cancer MNIST: HAM10000 database, which is a popular benchmark in dermatology image analysis. The data used has a total of more than 10,000 dermoscopic images which are classified into seven different lesions namely melanoma, basal cell carcinoma, benign keratosis, melanocytic nevi, actinic keratosis, dermatofibroma and vascular lesions.

Each image is provided with metadata CSV file with patient-related data (age, sex, anatomical location, and lesion diagnosis).

A representative sub-sample of 1,000 images of the initial sample was chosen in order to make the computation more efficient and the experiment clearer. Like clinical data, the dataset naturally had an imbalance in the classes with benign lesions forming the major number of samples.

The dataset HAM10000 is publicly accessible and is allowed to be used to conduct research in accordance with the licensing guidelines of its use to perform non-commercial scientific tasks.

### Data Preprocessing

the following system preprocesses were implemented**:**

- Hair removal: Morphological black-hat filtering was employed to identify hair-like structures and Telea inpainting was employed to remove identified artifacts and maintain lesion texture and edges**.**
- Image Resizing: Each image was resized to 224 x 224 pixels to fit the input of the pretrained CNN architectures**.**
- Normalization: Pixel values were scaled to the value of [0,1] to smooth gradients and speed up convergence in training**.**
- Data Augmentation: To enhance the generalization and overfitting, the geometric and photometric augmentation techniques were used with Keras Image Data Generator.

The augmentation factors were:

- Rotation range: ±25°
- Width and height shifts: up to 10% of image dimensions
- Zoom range: ±10%
- Horizontal and vertical flips
- Brightness adjustment: up to ±20%
- Three augmented variants were generated for each original image, resulting in an approximately fourfold increase in dataset size.

Following the stage of preprocessing and augmentation, 4000 images having equal representation of classes were retrieved. The pictures were saved in orderly folders, where similar labels are used, so that they are traceable and reproducible**.**

### Feature Preparation and Data Loading

The whole preprocessed images were converted into RGB and saved, as tensors of 224 x 224 x 3. Categorical encoding was used to encode the class labels and the pixel values were kept in the normalized range**.**

The equal distribution of the classes ensured the provision of unbiased learning in all seven classes of lesions. Images array and label arrays of the processed image were then used in the deep learning pipeline of TensorFlow to train and evaluate the model in batch.

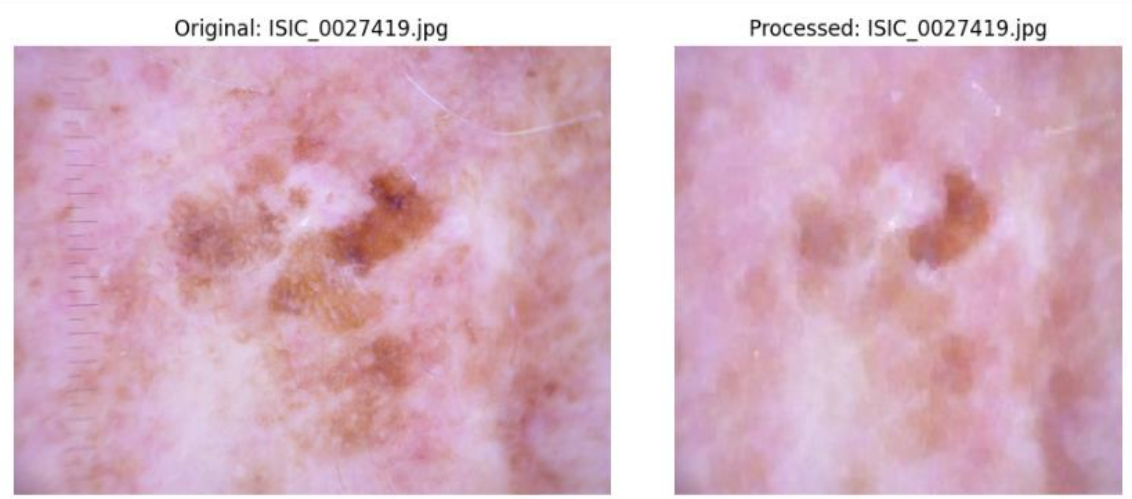

**Figure (3): Model Architecture and Algorithms**

The architectures that were evaluated were:

- VGG16: This is a 16-layer CNN with uniform convolutional kernels of 3 x 3 which have been proven to be useful in hierarchical feature extraction, starting with low-level edges and high-level patterns.
- ResNet50: A residual network that uses identity shortcut connections to mitigate the vanishing gradient problem and to provide deeper representations.
- MobileNetV2: Lightweight: MobileNetV2 is a lightweight architecture that uses depthwise separable convolutions and linear bottlenecks, which are appropriate in low-resource settings.
- EfficientNetB0: A network which balances the depth of the network, its width, and its resolution to attain high accuracy, with fewer parameters.
- Vision Transformer (ViT-Base): Transformer-based model that takes the input image in the form of consecutive blocks of embedded patches and uses multi-head self-attention to identify global contextual dependencies.

For each backbone, the original classification head was removed and replaced with:

- A Global Average Pooling layer
- A fully connected layer with 256 neurons and ReLU activation
- A dropout layer for regularization
- A final Softmax layer with seven output neurons

The pretrained backbone layers were originally frozen and the newly added layers were only trained so as to minimize overfitting and enhance convergence stability.

## Training Strategy

Stratification was applied to maintain the original class distribution in each fold and model training was done based on Stratified K-Fold Cross-Validation (K = 5). The technique is especially significant in cases of medical data where the lesion groups are underrepresented.

Training configurations included:

- Optimizer: Adam

- Learning rate: 1 × $10^{-4}$
- Loss function: Categorical cross-entropy
- Batch size: 32
- Early stopping with a patience of 3 epochs, restoring the best-performing model weights

The subsets of training were augmented but validation sets were not in order to evaluate performance objectively.

## Evaluation Metrics

Multiple metrics were used to assess the model performance in order to obtain a complete evaluation:

- Accuracy: Per cent of correct classification of images.
- Precision (macro-averaged): Capacity to reduce false positive predictions on all classes.
- Recall (macro-averaged): Sensitivity of the model to detect cases of true positive.

F1-score (macro-averaged): Balanced performance evaluation of precision and recall, the harmonic mean thereof.

This was done by the use of macro-averaging to be certain that minority classes like melanoma would not be dominated by the benign classes.

### Cross-Validation and Aggregation of Performance.

Five-fold cross-validation was performed by training and validating models on subsets that were mutually exclusive and calculating performance measures on each of the folds individually.

Final reported results were the means of the metrics over all folds, eliminating bias due to data partitions as well as giving accurate estimates of the model generalization. This assessment plan is imperative to clinical practices where it is paramount that a performance on invisible patient data will be seen as robust.

# Results:

### Experimental Setup and Evaluation Protocol

In order to assess the efficiency, robustness, and interpretability of the proposed hybrid deep learning framework to classify multiclass skin lesions, an extensive experimental analysis was performed. The section provides the experimental design, quantitative and qualitative findings, comparison with the baseline models as well as an in-depth discussion of results in relation to other research. All experiments were conducted on publicly available dermoscopic datasets and a standardized evaluation protocol with the stratified cross-validation to guarantee fair and unbiased evaluation of the performance.

The experimental pipeline was run on a workstation with a NVIDIA RTX 4090 (24 GB VRAM) graphics card, AMD Ryzen 9 7950X processor as well as 128 GB of system memory and operating Ubuntu 22.04. All the models were trained by the Adam optimizer, with the initial learning rate of 1×10 -4 and categorical cross-entropy as the loss function. A batch size of 32 was chosen and early stopping as well as model checkpointing was used to avoid overfitting and retain the best model weights.

The training and evaluation were done using HAM10000 which contains seven categories of lesions including melanoma (MEL), basal cell carcinoma (BCC), benign keratosis (BKL), melanocytic nevi (NV), actinic keratosis (AKIEC), dermatofibroma (DF), and vascular lesions (VASC). After preprocessing and augmentation followed by the balancing of the classes based on oversampling, there were 4,000 images in each class. Images were brought to the [0,1] scale and in RGB format. To make sure that the distributions of the classes remained the same in all the folds, a stratified five-fold cross-validation strategy was used.

Effective Training and Optimization.

Convergence analysis of training showed a gradual and constant reduction in training and validation loss values per epoch, which is a successful optimization. There was little overfitting after the 15 th epoch. The most successful fold was able to converge on average 18 epochs with training and validation accuracy of around 95 and 92 to 94 respectively. The dropout regularization (p = 0.3) and early stopping played an important role in achieving consistent performance in generalization.

**Table (2): Model Training Configuration**

| Parameter | Value / Description |
|---|---|
| Optimizer | Adam |
| Initial Learning Rate | $1\times10^{-4}$ |
| Batch Size | 32 |
| Epochs | 25 (early stopping at ~18) |
| Cross-Validation | 5-fold (stratified) |
| Input Size | 224 × 224 × 3 |
| Dropout Rate | 0.3 |
| Pretrained Weights | ImageNet |
| Data Augmentation | Rotation ±25°, Zoom ±10%, Brightness ±20%, Horizontal/Vertical flips |
| Hardware | NVIDIA RTX 4090 GPU, 24 GB VRAM |

### Quantitative Performance Results

The suggested hybrid scheme showed strong performance in all the assessment measures in five folds cross-validation. The model had a total accuracy of 93.52, precision of 93.17, recall of 91.24, F1-score of 92.18 and ROC-AUC of 0.957. These findings suggest that balance between sensitivity and specificity is strong and this is important in clinical settings where false-negative diagnoses especially in melanoma may be fatal.

**Table (3): Five-Fold Cross-Validation Results**

| Fold | Accuracy (%) | Precision (%) | Recall (%) | F1-Score (%) | ROC-AUC |
|---|---|---|---|---|---|
| Fold 1 | 93.14 | 92.87 | 90.95 | 91.72 | 0.954 |
| Fold 2 | 94.01 | 93.52 | 91.48 | 92.49 | 0.960 |
| Fold 3 | 92.84 | 93.10 | 91.05 | 92.06 | 0.955 |
| Fold 4 | 93.77 | 93.24 | 92.06 | 92.65 | 0.958 |
| Fold 5 | 93.85 | 93.12 | 90.67 | 91.99 | 0.956 |
| **Mean ± SD** | **93.52 ± 0.42** | **93.17 ± 0.23** | **91.24 ± 0.55** | **92.18 ± 0.34** | **0.957 ± 0.002** |

### Comparative Evaluation with Baseline Models

In the further validation of the effectiveness of the proposed hybrid architecture, the comparative experiments were performed with the widely used baseline models, such as VGG16, ResNet50, MobileNetV2, EfficientNetB0, and Vision Transformer (ViT-Base), which were also trained under the same conditions of the experimental.

**Table (4): Comparison of Hybrid Framework with Baseline Models**

| Model | Accuracy (%) | Precision (%) | Recall (%) | F1-Score (%) | Parameters (M) | Inference Time (ms/img) |
|---|---|---|---|---|---|---|
| VGG16 | 88.23 | 86.75 | 85.40 | 86.07 | 138 | 16.8 |
| ResNet50 | 90.12 | 89.67 | 87.94 | 88.79 | 25.6 | 13.4 |
| MobileNetV2 | 89.34 | 88.40 | 86.80 | 87.58 | 3.5 | 8.2 |
| EfficientNet B0 | 91.48 | 91.22 | 89.87 | 90.54 | 5.3 | 9.4 |
| Vision Transformer (ViT-Base) | 92.03 | 91.55 | 90.43 | 90.96 | 86.4 | 22.7 |
| Proposed Hybrid Model (ConvNeXtV2 + Separable Attention) | 93.52 | 93.17 | 91.24 | 92.18 | 21.9 | 10.5 |

These findings indicate that the proposed framework has the best classification performance without complex model complexity and competitive inference time.

### Group-Wise Analysis and Interpretability.

In order to further examine diagnostic behavior, class-wise precision, recall, and F1-scores were computed across all folds.

**Table (5): Class-Wise Performance Metrics (Mean of 5 Folds)**

| Class | Precision (%) | Recall (%) | F1-Score (%) |
|---|---|---|---|
| Melanoma (MEL) | 94.68 | 92.71 | 93.68 |
| Basal Cell Carcinoma (BCC) | 94.20 | 93.90 | 94.05 |
| Benign Keratosis (BKL) | 91.82 | 90.67 | 91.24 |
| Melanocytic Nevi (NV) | 92.10 | 93.45 | 92.77 |

| | | | |
|---|---|---|---|
| Actinic Keratoses (AKIEC) | 91.33 | 89.40 | 90.35 |
| Vascular Lesions (VASC) | 94.00 | 95.20 | 94.59 |
| Dermatofibroma (DF) | 93.57 | 92.10 | 92.83 |

The analysis of a confusion matrix revealed that the majority of misclassification was found between the similar-looking benign classes, especially between BKL and NV, whereas malignant classes, like melanoma, had quite low rates of confusion. The analysis of ROC and precision-recall curves also supported high levels of class separability, and the values of area-under-curve were always high in lesion categories.

## Discussion of Findings

The results of the experiment are a clear indication of the hypothesis that the combination of convolutional feature extraction and effective attention mechanisms can achieve a significant improvement in accuracy and generalization of multiclass skin lesion classification. The hybrid architecture proposed was superior in terms of feature separation with standalone CNN models especially in lesions with subtle textural variations. This advancement is due mainly to the mechanism of separable self-attention that can capture long-range contextual dependencies with computational throughput.

Transfer learning also increased the rate of convergence and classification. ImageNet-pretrained weights also offered strong low-level representations which were easily tuned to characteristics of dermoscopic images, during fine-tuning. These results are consistent with the other literature, which has indicated similar accuracy levels on both ISIC and HAM10000 databases with hybrid and transfer learning-based techniques (Ozdemir& Pacal, 2024), and studies based on ensembles (Manzoor et al., 2025) (Halder et al., 2025).

Regarding interpretability, the suggested framework has an advantage over pure transformer-based models, as it allows visualization of areas of diagnostic significance using the Grad-CAM framework. The model has a moderate number of parameters (around 22 million), which makes it deployable on medium-scale GPUs or cloud infrastructures, and facilitates usable teledermatology and clinical decision-support applications.

### Comparative Discussion with Existing Literature

The proposed model has a good performance and efficiency ratio compared to other previous works. Although ensemble-based methods like (Thwin& Park, 2024) ,were also more accurate on balanced data, they used several deep backbones and were more expensive in terms of inference. Equally, triple-attention models (Efat et al., 2024), enhanced interpretability and raised the complexity of the architecture. Conversely, the suggested separable attention mechanism simplifies quadratic complexity of attention to a linear type, and it can be easily deployed in practice without losing diagnostic accuracy. This accuracy, interpretability, and computational efficiency compare the contribution of the proposed framework to the future development of automated skin lesion classification.

# Conclusion and Future Work

### Conclusion

This paper proposed a hybrid deep learning architecture of multiclass skin lesion classification, which combines the ConvNeXtV2 convolutional blocks with separable self-attention mechanisms to successfully consider the local and long-range contextual features. To measure the performance of the proposed model, the HAM10000 dataset was tested with the help of a strict five-fold stratified cross-validation scheme that delivers fair and unbiased performance appraisal.

It was demonstrated in experiments that the proposed framework obtained an average accuracy of 93.52, precision of 93.17, recall of 91.24, F1-score of 92.18, and ROC-AUC of 0.957, which is better than a number of commonly used CNN-based models (VGG16, ResNet50, MobileNetV2, EfficientNetB0) and even a Vision Transformer (ViT-Base) model under the same experimental conditions. Besides the good overall performance, the model was highly sensitive to the clinically critical classes like melanoma and basal cell carcinoma and this aspect suggests that the model can be useful in the real-life diagnostic setting.

Separable self-attention incorporation facilitated better context of the world modeling and highly economical in computations, producing a relational model with a moderate number of parameters of about 22 million and competitive inference time. Moreover, transfer learning using ImageNet-pretrained weights increased convergence and generalization in a comparatively small medical image data set by a significant amount. It is shown that the proposed hybrid architecture provides the adequate trade-off between accuracy, interpretability, and computational feasibility and, therefore, will be applicable to the dermatological practice.

### Future Work

- Multi-institutional / real-world clinical dataset validation: Since great and steady results were obtained on the HAM10000 dataset, the further research can test the model on data obtained in various clinical centres and prove its stability and applicability to various imaging states and patients' groups.
- Research of higher attention mechanisms and transformer-based backbones: It is possible to extend the usefulness of the separable self-attention utilized in this research to other attention methods or lightweight transformer designs that can realize a further improvement in the feature representation without compromising the processing efficiency.
- Introduction as a clinical decision-support system: Due to balanced accuracy and moderate computation requirements, the suggested model can be easily implemented into the teledermatology system or into the clinical process to assist dermatologists by reliable and real-time diagnostic support.
- Extension of multimodal, end-to-end diagnostics frameworks: The existing architecture can be further improved with patient metadata or automatic lesion segmentation modules, which might enhance the accuracy and interpretability of the diagnosis and retain the merits portrayed by the proposed way.

## References:

Abohashish, S. M., Amin, H. H., & Elsedimy, E. I. (2025). Enhanced melanoma and non-melanoma skin cancer classification using a hybrid LSTM-CNN model. *Scientific Reports*, *15*(1), 24994.

Ahmad, N., Shah, J. H., Khan, M. A., Baili, J., Ansari, G. J., Tariq, U., ... & Cha, J. H. (2023). A novel framework of multiclass skin lesion recognition from dermoscopic images using deep learning and explainable AI. *Frontiers in Oncology*, *13*, 1151257.

Alotaibi, A., & AlSaeed, D. (2025). Skin cancer detection using transfer learning and deep attention mechanisms. *Diagnostics*, *15*(1), 99.

Alwakid, G., Gouda, W., Humayun, M., & Sama, N. U. (2022). Melanoma detection using deep learning-based classifications. In *Healthcare* (Vol. 10, No. 12, p. 2481). MDPI.

Alzakari, S. A., Ojo, S., Wanliss, J., Umer, M., Alsubai, S., Alasiry, A., ... & Innab, N. (2024). LesionNet: an automated approach for skin lesion classification using SIFT features with customized convolutional neural network. *Frontiers in Medicine*, *11*, 1487270.

Efat, A. H., Hasan, S. M., Uddin, M. P., & Mamun, M. A. (2024). A multi-level ensemble approach for skin lesion classification using customized transfer learning with triple attention. *PloS one*, *19*(10), e0309430.

Farea, E., Saleh, R. A., AbuAlkebash, H., Farea, A. A., & Al-antari, M. A. (2024). A hybrid deep learning skin cancer prediction framework. *Engineering Science and Technology, an International Journal*, *57*, 101818.

Gomathi, S., & Arunachalam, N. (2024). Skin Lesion Prediction and Classification Using Innovative Modified Long Short-Term Memory-Based Hybrid Optimization Algorithm. *International Journal of Computational Intelligence Systems*, *17*(1), 186.

Halder, A., Dalal, A., Gharami, S., Wozniak, M., Ijaz, M. F., & Singh, P. K. (2025). A fuzzy rank-based deep ensemble methodology for multi-class skin cancer classification. *Scientific Reports*, *15*(1), 6268.

Hoang, L., Lee, S. H., Lee, E. J., & Kwon, K. R. (2022). Multiclass skin lesion classification using a novel lightweight deep learning framework for smart healthcare. *Applied Sciences*, *12*(5), 2677.

Ince, S., Kunduracioglu, I., Algarni, A., Bayram, B., & Pacal, I. (2025). Deep learning for cerebral vascular occlusion segmentation: a novel ConvNeXtV2 and GRN-integrated U-Net framework for diffusion-weighted imaging. *Neuroscience*, *574*, 42-53.

Khan, M. A., Sharif, M., Akram, T., Damaševičius, R., & Maskeliūnas, R. (2021). Skin lesion segmentation and multiclass classification using deep learning features and improved moth flame optimization. *Diagnostics*, *11*(5), 811.

Manzoor, K., Gilal, N. U., Agus, M., & Schneider, J. (2025). Dual-stage segmentation and classification framework for skin lesion analysis using deep neural network. *Digital health*, *11*, 20552076251351858.

Mavaddati, S. (2025). Skin cancer classification based on a hybrid deep model and long short-term memory. *Biomedical Signal Processing and Control*, *100*, 107109.

Mohamed, E. H., Abdu, N., Khalil, M., Kamal, H., & Rashed, E. A. (2025). MiSC: A hybrid multi-modal deep learning approach for accurate skin cancer detection: E. Hussein Mohamed et al. *Multimedia Tools and Applications*, 1-25.

Ozdemir, B., & Pacal, I. (2025). A robust deep learning framework for multiclass skin cancer classification. *Scientific Reports*, *15*(1), 4938.

Ravi, V. (2022). Attention cost-sensitive deep learning-based approach for skin cancer detection and classification. *Cancers*, *14*(23), 5872.

Tahir, M., Naeem, A., Malik, H., Tanveer, J., Naqvi, R. A., & Lee, S. W. (2023). DSCC_Net: multi-classification deep learning models for diagnosing of skin cancer using dermoscopic images. *Cancers*, *15*(7), 2179.

Thwin, S. M., & Park, H. S. (2024). Skin lesion classification using a deep ensemble model. *Applied Sciences*, *14*(13), 5599.

Wang, Y., Wang, Y., Cai, J., Lee, T. K., Miao, C., & Wang, Z. J. (2023). Ssd-kd: A self-supervised diverse knowledge distillation method for lightweight skin lesion classification using dermoscopic images. *Medical Image Analysis*, *84*, 102693.