JCBI
Journal of Computing & Biomedical Informatics
ISSN: 2710 - 1606
Volume 11 Issue 02
2026

Review Article
https://doi.org/10.56979/1102/2026/1411

Automated Skin Cancer Classification from Dermoscopic Images Using Fine-Tuned CNN Models

Muhammad Shoaib Rasheed[1], Muhammad Asim Rajwana[1], Sadia Tariq[1], Tahir Abbas[2*], Nazir Ahmad[1], Zaiba Aziz[1], and Ehsan ul Haq[1]

[1]Department of Computer Science, National College of Business Administration & Economics Lahore, Multan Sub Campus, 60000, Pakistan.
[2]Department of Communication and Cyber Security, Bahauddin Zakariya University, Multan, 60000, Pakistan.
[*]Corresponding Author: Tahir Abbas. Email:tahir.abbas@bzu.edu.pk

Received: March 23, 2026 **Accepted:** July 04, 2026

Abstract: Skin cancer mostly melanoma cause a significant health risk as the dieses spread quickly and high risk of death if not detected early. The early detection of skin disease is essential but depending only on visual examination can be difficult for dermatologists and it may lead to delayed diagnoses. The deep learning based diagnostic systems can support clinicians by classifying skin lesions in a consistent manner and reducing variability in visual assessment. To automatically classify dermoscopic images we introduce a machine learning model that combines transfer learning with pre-trained CNNs. Three architectures VGG16, MobileNetV2 and EfficientNetB0 were evaluated on the MED-NODE dataset which contains images of melanoma and naevus that limited and imbalance. The methodology integrates the methods such as image pre-processing, extensive data augmentation, selective fine-tuning, hierarchical feature extraction and class weighting to improve model generalization and handle dataset limitations. The experimental results show that VGG16 achieved the highest performance with 82.5% accuracy, 84.67% weighted precision, 82.22% F1-score and 91.25% AUC. MobileNetV2 achieved moderate performance, whereas EfficientNetB0 performed poorly. This highlights the importance of deeper pretrained features and careful adaptation for small and imbalanced medical datasets. Analysis of per class performance, confusion matrices and training/validation curves shows that VGG16 extracts low-level features including edges, textures and also identifies higher-level patterns like asymmetry, borders and lesion shape which helps distinguish melanoma from naevus more reliably. The results show that transfer learning with deep CNN models improves automated skin lesion classification. The performance further improves with data pre-processing, data augmentation, class weighting and fine-tuning strategies. We proposed a machine learning framework for melanoma detection that helps clinicians to achieve more reliable diagnostic results in clinical practice. The framework can be evaluated on larger datasets using hybrid models to improve both performance and interpretability.

Keywords: Melanoma; Skin Cancer; Dermoscopic Images; CNN; Transfer Learning; VGG16; MobileNetV2; EfficientNetB0; Data Augmentation; Imbalance Dataset

1. Introduction

Globally skin cancer is one of the leading health concerns and timely diagnosis plays an important role to reduce its impact. Some of skin cancers are consider to be serious because they can spread rapidly to the other parts of body the melanoma is one of them. The appearance of unusual moles or spots on the skin should be examined as early as possible. The early detection and treatment can save human lives [1]. Traditionally, dermatologists examine skin lesions through visual inspection and by using diagnostic tools such as dermoscopy to identify suspicious patterns [2, 3]. However, many skin lesions have only small differences in color, texture or shape. These differences make visual diagnosis tough even for qualified

clinicians [4]. The diagnostic process depends on the experienced and trained specialists and also require significant time. In many regions specialists are not easily available which further complicates early diagnosis [5]. These limitations initiate a need for deep learning systems that can assist clinicians in the detection and classification of skin lesions [6]. Deep learning models have shown strong potential to support medical professionals by providing consistent and accurate analysis of medical images [7, 8]. The recent advancement in deep learning has introduced effective methods for medical image analysis [9]. CNNs are widely used for image based medical diagnosis because they can learn hierarchical features directly from images [10]. In a CNN the first layers detect basic image features such as edges and texture patterns. The deeper layers capture more complex information including lesion shapes, borders and structural patterns [11]. Transfer learning further improves performance by using networks that were previously trained on large scale datasets and adapting them to smaller medical image datasets. This method reduces the training time and also improves classification accuracy [3]. Several well-known CNN architectures such as VGG16, MobileNetV2 and EfficientNetB0 have produced strong results in image classification tasks and have also shown promising performance in medical image analysis [12].

Even with recent development the machine analysis of skin lesions still faces major challenges. Many studies focus on a single CNN architecture and limiting understanding of which models perform best [13]. Medical imaging datasets are often small and imbalanced which can lead to overfitting and reduce model generalization. Lesions from different classes may appear visually similar that is making it difficult for models to learn minor discriminative features [14]. A systematic evaluation of multiple CNN architectures under consistent experimental conditions is therefore necessary to determine the highly effective method for skin lesion classification. This study uses the MED-NODE dataset which contains of dermoscopic images of melanoma and naevus [3, 15]. The dataset is small, imbalanced and contains visually similar lesions which increases the difficulty of learning robust feature representations. To address these issues methods like pre-processing, data augmentation and class weighting are implemented to enhance model training and generalization [16]. We perform a comparative analysis of three transfer learning-based CNN architectures for automated skin lesion classification. All models are pretrained, fine-tuned and evaluated on the same benchmark dataset using standardized training procedures. With this global contextual information captured transformer-based and hybrid deep learning models are recently a topic of interest in medical image analysis. In small medical data sets, such as MED-NODE they are often less successful, due to the necessity of large data sets.

1.1. Recent Advances in Medical Image Analysis and Explainable AI

More recently, developments with transformers and their combination with convolutional neural networks (CNNs) in hybrid architectures for deep learning have significantly pushed the boundaries for analysis of medical images by incorporating attention mechanisms for modelling long-range context, proving effective for large-scale image classification such as in skin lesion classification. In tandem, interest in explainable artificial intelligence (XAI) techniques such as Grad-CAM and SHAP [17] has risen as models benefit from having visual and feature representations to aid the clinical trust and utilization process. There are suggestions that for reliable and robust deployment of AI models within medical workflows, excellent predictive accuracy and robust interpretability must be simultaneously presented. However, many of these modern techniques demand abundant annotation information and heavy computation, posing a barrier when working with scarce and imbalanced dataset like MED-NODE[18]. Hence, efficient and interpretable methods in deep learning that yield satisfactory performance on low-resource conditions are required.

1.2. Main Contributions of This Study

This work aims at solving the problems concerning skin lesion classification methods based on CNNs which are limited to limited size, imbalanced datasets and inconsistent evaluation pipeline. The main contributions of this work are:

1. United and Controlled Benchmark Framework. This work is the first that ensures a fair and consistent comparison among VGG16, MobileNetV2 and EfficientNetB0 on MED-NODE dataset with identically the same input pre-processing, augmentation, and training strategies applied.

2. Advanced Deep Learning Algorithms in Big Data Scenario. Hybrid strategy leveraging data argumentation, class weighing, and partial fine-tuning is proposed and demonstrated as a well-suited strategy for data-scarce medical imaging datasets to achieve better generalization[19].
3. Understanding the Feature Learning Behaviour. Layer-wise analysis for extracting features by the CNN is conducted to elucidate that both the high-level features (i.e., asymmetry, structure of lesions) and low-level features (i.e., edges, textures) important for melanoma classification.
4. Assessment for Clinical Deployment. In spite of general belief, it is highlighting that deep architectural models over-perform lightweight ones in such data restricted situations and offer guidance on developing real world CAD systems.

Although a significant number of CNN-based solutions for skin lesion classification has been proposed so far, the existing literature has certain shortcomings. Firstly, few works examine these multiple CNN-based models, as the majority of studies focus on single architectures or adopt entirely different datasets and experimental settings which do not permit a balanced assessment. Secondly, with the exception of a few recent publications, most works perform an evaluation, but do not study any specific situation in which some important constraint applies, e.g., few and imbalanced data that are characteristic in a real clinical environment. Thirdly, the impact of some commonly used training strategies (e.g. Data augmentation, weighting and finetuning) in terms of generalization behaviour and the process of feature learning remains unexplored in a consolidated way. The aim of this study is to perform such study within a unique, controlled and exhaustive framework by applying numerous already-trained CNN architectures to a common skin dataset in homogeneous circumstances. More concretely, through a single and complete experimental campaign on MED-NODE, the influence of training on model robustness, learning process, specific features generated, various training aspects will be clearly analyze.

2. Related Work

[20] proposed an AI framework for classifying skin lesions as fundus or non-fundus during teledermatology screening using patient metadata and dermoscopic images. The study used 79246 skin lesion images with 22 meta features from 19295 patients where the metadata model gains 85.24% sensitivity and 61.12% specificity. The image model attain 99.72% sensitivity and 63.22% specificity. The fused model hits 99.66% sensitivity and 74.45% specificity while the majority voting improved performance to 99.50% sensitivity and 82.72% specificity.

[21] proposed a novel method for classifying skin lesions by combining deep learning with optimization techniques. The model uses RegNetY032 with a modified classification head and a Soft Attention Block to focus on salient lesion features while ignoring artifacts and employs Harris a proposed model that achieved a classification accuracy of 99.27% demonstrating its potential for improving automated dermatological diagnosis and early skin cancer intervention.

[22] conducted a benchmarking study of 38 CNN architectures of ten families for automated seven classes skin lesion classification using transfer learning on the HAM10000 dataset. The Cross-database validation with ISIC2019 assessed generalizability with ConvNeXtXLarge achieving the highest performance of 87.62% accuracy and 76.15% F1 score. A preliminary study of mobile application demonstrated the feasibility of implementing high performing models in healthcare screening tools.

[23] design a diagnostic model that was capable of classifying multiple skin lesions using CSMUH dataset that was focusing on the population from east Asia. Twenty-five pre-trained CNN and ViT models were refined and the top three were combined to form ensemble Swin-ViT-EfficientNetB4 with hard and soft voting strategies. The model was evaluated using five randomized experiments with holdout validation. The ensemble achieved a test accuracy of 98.5% that demonstrating the ability to accurate and early skin lesion diagnosis.

[24] proposed DermViT that a medically inspired deep learning architecture for skin lesion classification that replicate physician diagnostic processes. DermViT includes the Dermoscopic Context Pyramid (DCP) for multi-scale feature extraction, Dermoscopic Hierarchical Attention (DHA) for focused lesion analysis and Dermoscopic Feature Gate (DFG) to suppress artifact interference. Evaluated on ISIC2018 and ISIC2019 datasets, DermViT achieved 86.12% accuracy that was a 7.8% improvement over ViT-Base with 40% fewer parameters that demonstrating efficient and accurate analysis of dermoscopic image.

[25] developed a deep learning web application for skin lesion classification and analysis. Various DL models were evaluated from them the DenseNet169 achieving the highest accuracy of 85% and selected as the base model. The platform provides image classification with confidence scores, visualizes convolutional features for interpretability and delivers evidence-based dermatological information that supporting early diagnosis and public education.

[26] introduced a model of machine learning for multi-class skin lesion classification using HAM10000. CNNs and Vision Transformers were evaluated with InceptionResNetV2 achieving 88% accuracy and 77% F1-score. A weighted soft voting ensemble improved performance to 89% accuracy and 80% F1-score that was highlighting the effectiveness of transfer learning and ensemble methods.

[27] proposed MaLafFormer that was a Swin Transformer model for dermoscopic skin lesion classification that combines a MFMA module and a LAF module to capture both global and local features. The proposed model was tested on the ISIC2018 dataset where MaLafFormer achieved 84.35% accuracy, outperforming the baseline by 6.37% and surpassing other compared methods, demonstrating its effectiveness for skin lesion analysis.

[28] proposed a federated ViT framework for skin lesion classification using HAM10000 and ISIC2019 datasets. The model achieved 88.2% sensitivity and 91.4% specificity on HAM10000 and 87.6% accuracy with 96% AUC on ISIC2019 that demonstrating strong performance on heterogeneous and privacy-sensitive data.

3. Materials and Methods

We present a deep learning method for classification of skin lesions using dermoscopic images. The pre-processing and augmenting methods are used on the dataset to make the model more efficient. Then we apply transfer learning using a pre-trained CNN and then we fine-tune them to achieve the specific task. The classification is improved by extracting hierarchical features that enable the model to manage small and imbalanced dataset to maintain consistent performance.

3.1. Dataset Description and Pre-processing

This study the has been based on MED-NODE data consist of total 170 dermoscopic images[15]. The dataset consists of 70 melanomas and 100 naevus. It can be inferred, that this dataset is relatively small class imbalanced that means naevus class greater than melanoma class. The fact that might skew the results towards the majority class and thus will affect negatively the classification performance. The Class distribution can be illustrated via Table 1. The trained models have been evaluated on validation set, for all stated measures. To make the data not leak to unseen samples for fairness assessment and train the model the training set the data augmented. Images were resized to 224 X 224 pixels which is size of pre-trained CNN input size and for helping the model during training convergence, pixels were scaled to the [0,1] rage of integer value[29]. By using data augmentation techniques including the random rotations between -20 +20 degrees random vertical and horizontal flips, zooming between 0.8 to 1.2, width shifting between -0.2 to 0.2, height shifting between -0.2 to 0.2 and brightness change between-0.2 to 0.2. Thus, all above data augmentation schemes enable model to learn the invariant features which are discriminatory in the presence of scarce samples. Class weighting was also considered in the models training making it to place relatively more weight on minority classes[30]. Thus, the class imbalanced bias was also reduced to its minimum level and the performance measures on melanomas would be increased. Due to limited data and variety in them, generalization was a challenge; Although, in this study MED-NODE dataset was augmented (images) for diversity as well as classes were weighted to overcome class imbalance, yet it was small dataset that causes a poor performance for prediction across the entire population and more generally for melanoma class prediction. The data contained mainly only two cases thus do not portray the broad range of circumstances seen with typical clinical applications (skin diversity, varying photo environmental factors, etc). Therefore, the proposed model might not demonstrate similar behaviour to broader sets of larger data. Future scope of this work is validation of the current architecture to a greater dataset such as IISC along with increasing diversity to achieve a robustness in the clinical scope with even greater range of data diversity. Validation Split of 80:20 split between the training set and validation set was applied by using the native validation parameter of the data generator. The split at image level took place in a way that an image will not have a belonging of the both sets. Only a training data generator

applied Data Augmentation to enhance the number of samples and decrease over fitting the validation generator kept out for example this will apply Data Augmentation to no one among them only the test one and validation generator is using same samples no further augmented to test only with the original images. Therefore, no over fitting will happen. At the level of the selection of train or validate data validation generator used 'set'. To avoid leak through over fitting to each step validation set is totally independent. Data Shuffling is available on train set only.

Table 1. Class distribution of the MED-NODE dataset

Name of Dataset	Melanoma	Naevus
MED-NODE	70	100

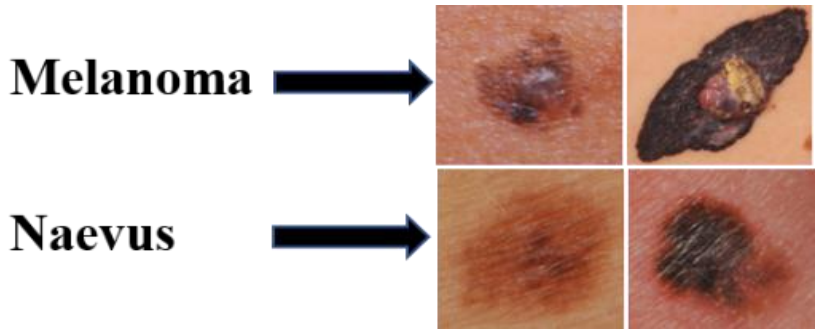

Figure 1. Data Set melanoma and naevus images

3.2. Model Architecture and Training

Three pre-trained convolutional neural network (CNN) models namely, VGG16, MobileNetV2 and EfficientNetB0 were used for automated skin lesion classification. They were all started with the ImageNet weights and adapted to binary classification by removing the fully connected layers and replacing them with a custom classification head such as Global Average Pooling (GAP) layer, a dense layer having 256 neurons with ReLU activation, a dropout layer with drop rate of 0.5 and a softmax output layer[31]. Training process was split into two stages to make sure of effective transfer learning. In the first step, all the convolutional base layers were frozen and only the added classification layers were trained for 15 epochs. The second step was to fine-tune the model with just the top layers unfrozen and the remaining layers frozen, and then train it again for 15 more epochs. The final 4 layers of VGG16 were unfrozen, as well as the final 20 layers of MobileNetV2 and EfficientNetB0. In the first stage, all the models were trained with the Adam optimizer using a learning rate of 1×10^{-4} and in the second stage, fine-tuned using a learning rate of 1×10^{-5} [32]. The batch size was set to 16 and the loss function was set to categorical cross-entropy. For training stability, and to avoid over fitting, early stopping (patience = 6 epochs) and learning rate reduction on plateau (factor = 0.2, min learning rate = 1×10^{-7}) were used according to the validation loss. Class weights were calculated and used in training to deal with class imbalance. All the models were tested with the same training configuration, pre-processing pipeline and evaluation protocol for a fair and reproducible comparison.

3.3. Convolutional Neural Network (CNN)

The proposed model uses VGG16 pretrained on ImageNet for hierarchical feature extraction [33]. VGG16 has 13 convolutional layers with 3 fully connected layers divided in five blocks that are extracting features ranging from simple patterns to complex high-level patterns. All convolutional layers use 3 × 3 kernels with stride one and ReLU activation. The max pooling reduces feature map size while preserving key activations.

$$g(z) = max(0, z), \qquad p_{u,v} = max\{z_{s,t} \text{ in pooling window}\} \tag{1}$$

The top layers are replaced with GAP), dense layer with ReLU activation, dropout layer and a softmax that allowing the model to trageted on lesion-relevant features.

3.4. Feature Representation and Hierarchical Feature Extraction

The CNNs automatically extract structured features by selecting low-level patterns such as edges, textures and high-level semantic features such as lesion shape, asymmetry and borders. The fine-tuning

maintains pretrained features while adapting to dataset-specific characteristics maintaining critical information that are necessary for small datasets. This improves generalization and ensures discriminative lesion characteristics are retained.

3.5. Performance Evaluation

The performance evaluation metrics are used to evaluate models performance that were derived from the confusion matrix. The experiments were repeated for R independent runs where r=1, 2, …, R.

For the r^{th} run the dataset distribution matrix is

$$H^r = \begin{bmatrix} \frac{|D_{MEL}|}{2} & 0 \\ 0 & \frac{|D_{NAE}|}{2} \end{bmatrix} \quad (2)$$

where

- $|D_{MEL}|$ is the total number of melanoma images.
- $|D_{NAE}|$ is the total number of naevus images.
- For balanced sampling class it is divided by 2.

The confusion matrix for the r^{th} run is shown below

$$M^r = \begin{bmatrix} t_r & f_r \\ m_r & n_r \end{bmatrix} \quad (3)$$

where

- t_r is true positive
- f_r is false positive
- m_r is false negative
- n_r is true negative

The performance metrics shown as

Accuracy

$$A^r = \frac{t_r + n_r}{t_r + f_r + m_r + n_r} \quad (4)$$

Precision

$$P^r = \frac{t_r}{t_r + f_r} \quad (5)$$

Sensitivity (Recall)

$$R^r = \frac{t_r}{t_r + m_r} \quad (6)$$

Specificity (True Negative Rate)

$$S^r = \frac{n_r}{n_r + f_r} \quad (7)$$

F1-score

$$F^r = \frac{2 \times P^r \times R^r}{P^r + R^r} \quad (8)$$

Area Under the ROC Curve is

$$U^r = \int_0^1 TPR(FPR)d(FPR) \quad (9)$$

Finally, the average metrics are

$$\bar{X}_i = \frac{1}{R}\sum_{r=1}^{R} X_i^r \ , \qquad X_i \in \{A, P, R, F, U\} \quad (10)$$

Where the TPR is the true positive rate and FPR is the false positive rate from the ROC curve. The confusion matrices were used for class-level assessment and training/validation curves used to check convergence, detect overfitting and compare the three architectures. Figure 2 represents the complete architecture of the proposed model.

An evaluation strategy based on validation was used to reduce the risk from overfitting caused by the small dataset. The data was divided into training and validation sets at random proportions of 80:20 using a data generator such as controlled random data split with a validation split parameter. The augmentation of the data was used only for the training set, to increase the variation and the generalisation of the model, and the validation set was not seen during the training. Regularization methods such as dropout, early stopping using the validation loss, and learning rate dropping were also used to avoid overfitting and train the model stably. To deal with the imbalance in the data set, class weighting was also included. The used validation approach and regularization methods are sufficient for the present limited set of data to give a reliable estimation of the model performance, Although the validation-based evaluation provides a reasonable estimate of model performance for the current dataset, more robust validation strategies such

as k-fold cross-validation or independent external testing would further improve reliability and generalization. These approaches will be considered in future work to strengthen the evaluation framework.

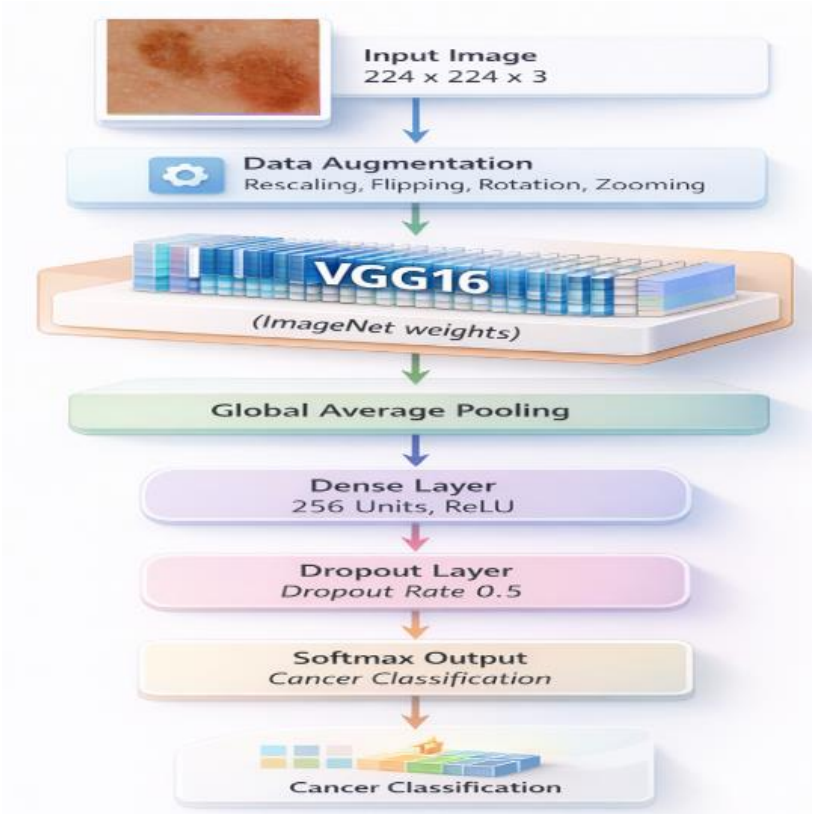

Figure 2. Proposed VGG16-Based Deep Learning Architecture for Cancer Image Classification.

3.6. Experimental Workflow

Figure 3 presents the overall workflow. After pre-processing and data augmentation the images are input into CNN architectures. The pre-trained networks are subsequently modified and fine-tuned using class weighting and hierarchical features are extracted for analysis. The model performance is then assessed using standard evaluation metrics.

3.6.1. Proposed Workflow for Automated Skin Lesion Classification

1. **Dataset Loading:** In this study the dermoscopic images of Melanoma and benign naevus lesions were loaded from the MED-NODE Dataset for analysis.
2. **Pre-processing:** All images were resized to **224 × 224 pixels** and the pixel intensity values were then normalized.
3. **Data Augmentation:** By applying data augmentation techniques such as rotations, shifts, zooming, flips, shearing and brightness adjustments to increase dataset diversity.
4. **Model Selection:** We use three CNN architectures such as VGG16, MobileNetV2 and EfficientNetB0 with pretrained ImageNet weights.
5. **Model Modification:** Replace top layers with a global average pooling layer, dense layer 256 units, ReLU, dropout layer (0.5) and softmax output layer for classification.
6. **Training:**

- **Stage 1:** Freeze base layers and then train new layers.
- **Stage 2:** Selectively fine-tune deeper layers.
- Apply class weights to address imbalance, use learning rate scheduling and early stopping.

7. **Feature Extraction:** Hierarchical features are automatically extracted and preserved to maintain pretrained representations while adapting to dataset-specific patterns.
8. **Performance Evaluation:** Assess models using accuracy, precision, recall, F1-score, AUC and confusion matrices. Monitor training and validation curves for convergence and overfitting.

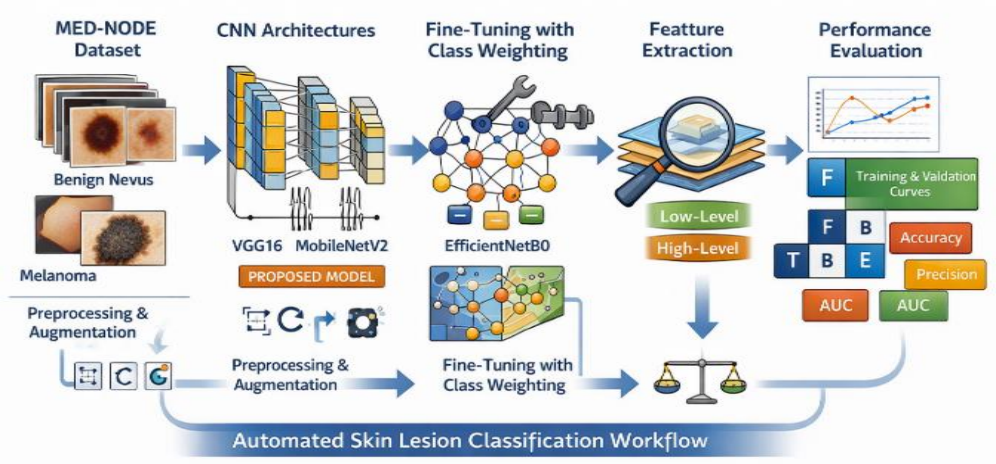

Figure 3. Workflow of the automated skin lesion classification system.

3.6.2. Transition to Results

Following model training and evaluation the performance of VGG16, MobileNetV2 and EfficientNetB0 was systematically compared using accuracy, precision, recall, F1-score and AUC. Confusion matrices and training and validation curves were analysed to assess class-level performance and convergence trends that providing a clear basis for identifying the most effective architecture for automated skin lesion classification.

4. Results

The proposed deep learning framework was evaluated on the MED-NODE dataset comparing proposed model, MobileNetV2 and EfficientNetB0. Model performance was assessed using standard classification metrics such as accuracy, precision, recall, F1-score and AUC. Table 2 summarizes the overall performance of each model.

Table 2. Overall Model Performance

Model	Accuracy%	Precision%	Recall%	F1-score%	AUC%
VGG16	82.5	84.67	82.5	82.22	91.25
MobileNetV2	77.5	78.13	77.5	77.37	86.25
EfficientNetB0	50	25	50	33.33	52

Among the all three architectures **VGG16 achieved the highest overall performance that** demonstrating the effectiveness of pretrained hierarchical features, selective fine-tuning and class weighting. MobileNetV2 achieved moderate performance while EfficientNetB0 performed poorly that is highlighting the network scaling alone is insufficient for small imbalanced dermoscopic datasets.

4.1. Per-Class Performance

Per-class metrics were analysed to assess model capability in distinguishing melanoma from naevus. The VGG16 correctly identified **melanoma samples with 82.5%** sensitivity **(recall)** that demonstrating its ability to handle the minority class despite dataset imbalance. Its weighted precision for melanoma was 84.67% indicating few false positives. The MobileNetV2 misclassified a higher proportion of melanoma images as benign recall 77.5 whereas EfficientNetB0 struggled with both classes that is reflecting insufficient feature extraction capacity for this dataset.

4.2. Confusion Matrices

Confusion matrices for VGG16 and MobileNetV2 are shown in Figure 3. VGG16 correctly classified the majority of both melanoma and benign samples including challenging minority cases, that is demonstrating the positive effect of hierarchical feature extraction and class weighting. MobileNetV2 incorrectly classified some melanoma lesions which may be related to its relatively shallow architecture. EfficientNetB0 was not included in the final analysis because its performance was lower than the other models.

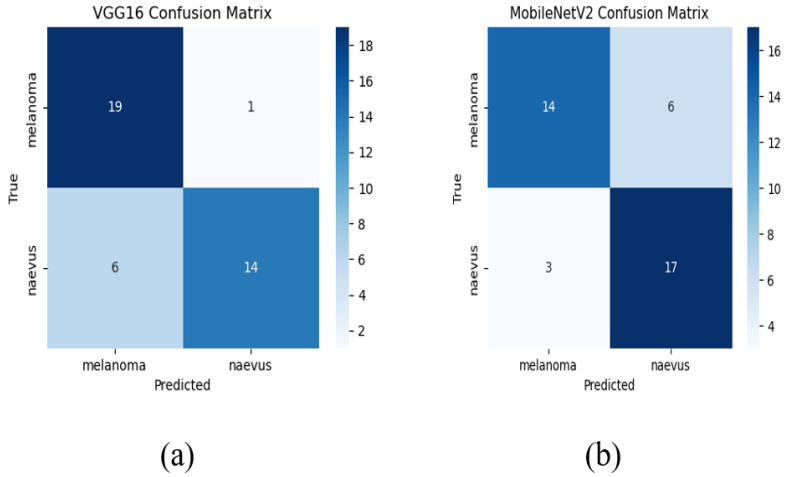

Figure 4. Confusion matrices (a) VGG16, (b) MobileNetV2.

4.3. Training and Validation Analysis

Figure 4 shows the training and validation curves. VGG16 converges within 30 epochs with training and validation accuracy remaining close and the loss curves decreasing smoothly that is suggesting good generalization. MobileNetV2 converges reasonably well though there are small fluctuations in validation loss. EfficientNetB0 on the other hand learns irregularly and struggles with feature extraction on small datasets.

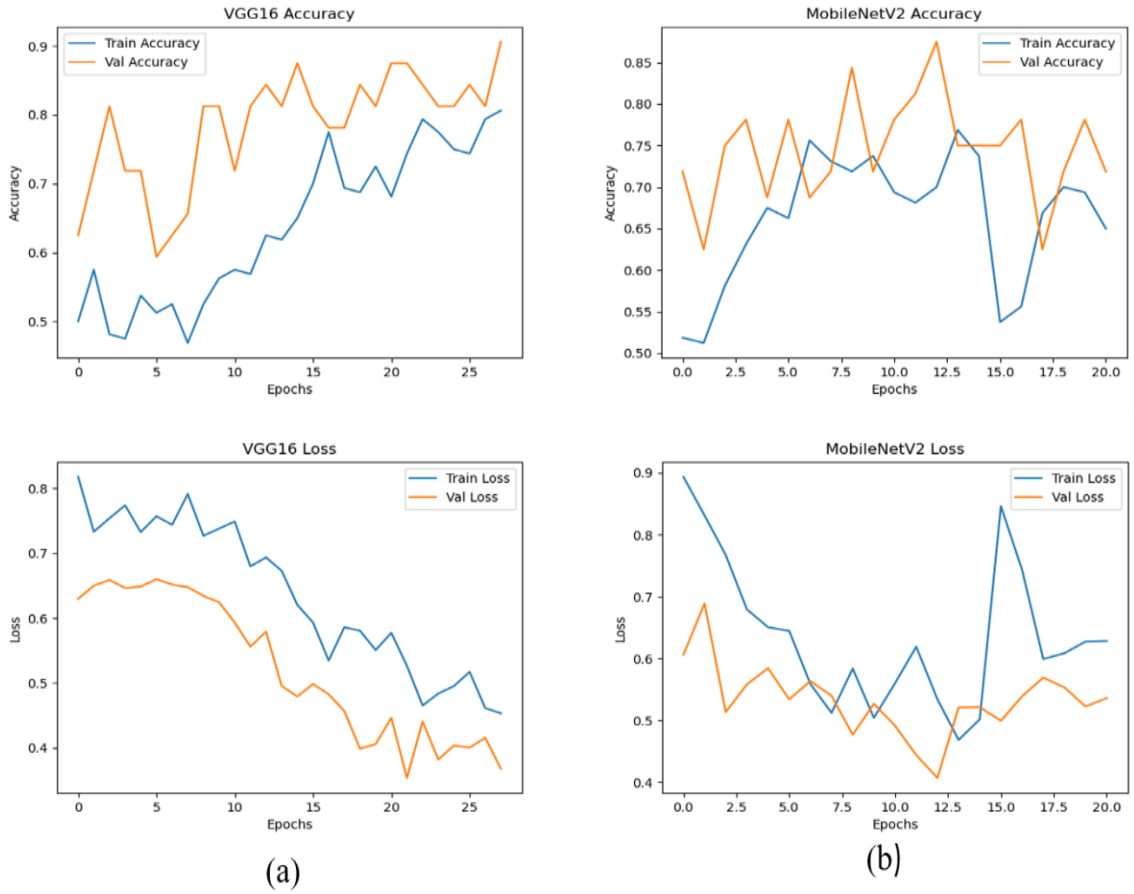

Figure 5. Training/validation accuracy and loss curves for the models (a) VGG16, (b) MobileNetV2.

4.4. Feature Analysis

VGG16's superior performance is attributed to its deeper convolutional layers which capture both low-level patterns edges, textures and high-level features asymmetry, lesion borders, shape. Fine-tuning

preserves pretrained ImageNet features while adapting to dataset-specific patterns. MobileNetV2 offers a trade-off between efficiency and accuracy while EfficientNetB0 fails to capture discriminative features without additional adaptation or fine-tuning. The results confirm that transfer learning with deeper CNN architectures combined with data augmentation, class weighting and selective fine-tuning significantly improves automated skin lesion classification**.** VGG16 consistently outperforms the other models across all metrics that highlighting the importance of hierarchical feature extraction and careful network adaptation for small and imbalanced dermoscopic datasets.

4.5. Validation Consistency

The reliability of the reported results was further assessed by analysing training and validation curves which demonstrate stable convergence behaviour across epochs, particularly for the VGG16 model. The close alignment between training and validation accuracy along with the smooth reduction in loss indicates minimal overfitting and consistent learning behaviour. Although repeated experimental runs and statistical significance testing were not performed due to dataset limitations, the consistency observed across multiple evaluation metrics and validation trends supports the robustness of the reported results.

5. Discussion

The evaluation of proposed model VGG16, MobileNetV2 and EfficientNetB0 on the MED-NODE dataset demonstrates that transfer learning with deeper CNN architectures significantly improves automated skin lesion classification. Proposed model consistently achieved the highest overall performance with 82.5% accuracy, 84.67% weighted precision, 82.22% F1-score and 91.25% AUC. This superior performance is primarily due to its deeper convolutional layers which extract rich hierarchical features encompassing both low-level patterns such as edges and textures and high-level semantic features like lesion asymmetry, borders and shape [34]. Fine-tuning of selective deeper layers preserved pretrained ImageNet representations while allowing adaptation to dataset-specific patterns that effectively capturing discriminative features critical for differentiating melanoma from naevus [35]. MobileNetV2 though more computationally efficient, exhibited moderate performance 77.5% accuracy and 86.25% AUC. Its shallower architecture and fewer convolutional filters limit the extraction of high-level lesion features that leading to misclassification of minority class melanoma images. These findings indicate that while lightweight architectures may be suitable for real-time applications and deeper networks offer better feature representation for small, imbalanced dermoscopic datasets [36].

EfficientNetB0 performed poorly 50% accuracy, 52% AUC, despite its compound scaling of depth, width and resolution. This means that simply scaling the model is not enough for small, imbalanced datasets unless it is combined with techniques like data augmentation, class weighting or selective fine-tuning [37]. The highly optimized model struggles on small medical datasets as reflected by the irregular learning in the curves. By augmenting the data with rotations, flips, shifts, zoom, shearing and brightness changes the models understood more diverse lesion patterns which improves generalization [38]. The performance of the model improves and it correctly identifies melanoma samples upto 82.5% recall by applying class weighting on imbalance medical dataset. This emphasizes how important it is to use careful pre-processing and training approaches [39]. Overall, these results show that by combining transfer learning with deeper pretrained CNNs, selective fine-tuning and hierarchical feature preservation is effective. In recent years, new techniques in medical image analysis have emerged such as transformer-based deep learning models and hybrid models like CNN–transformer, which have shown promising results in the field of skin lesion classification. Transformer-based deep learning models and hybrid models like CNN–transformer are recent techniques that have shown strong performance in medical image analysis tasks particularly in skin lesion classification [40]. The models utilize self-attention to encode the global context information of images and outperform the CNN-based models in image processing tasks on large-scale datasets. In particular, hybrid models that combine convolutional feature extraction with transformer-based attention provide an effective balance between local feature learning and global context modelling. But the transformer models are generally large-scale annotated datasets and high computational resource are needed for effective training. Their capability might be restricted in small and imbalanced datasets like MED-NODE if no extensive pretraining or data scaling are performed. However, transfer learning using CNN architectures is still a viable and efficient approach when data is limited. Thus,

in this study, models based on CNN are studied with the awareness that transformers and hybrid models were studied in the future to further enhance classification performance and interpretability. To improve the robustness and accuracy of skin lesion classification models it is necessary to properly handle dataset limitations [41]. VGG16 performs in a way that shows hierarchical feature extraction are more crucial than efficiency or scaling when dealing with small and imbalanced dermoscopic datasets.

5.1. Comparison with Recent Advanced Methods

Recent studies in skin lesion classification have explored advanced deep learning approaches, including transformer-based models and hybrid CNN–transformer architectures which have demonstrated strong performance on large-scale datasets such as ISIC [42]. These methods leverage self-attention mechanisms to capture global contextual information and have shown improved accuracy compared to traditional CNNs in data-rich environments. However, their performance is highly dependent on large annotated datasets and significant computational resources. In contrast, the proposed CNN-based framework achieves competitive performance on the relatively small and imbalanced MED-NODE dataset, with VGG16 reaching 82.5% accuracy and 91.25% AUC. Compared to recent CNN-based studies on small dermoscopic datasets, the results demonstrate that carefully fine-tuned deep CNN architectures combined with data augmentation and class balancing strategies can achieve reliable performance without requiring complex architectures. This highlights the practicality and effectiveness of the proposed approach in low-data clinical scenarios. Future work will include direct experimental comparisons with transformer-based and hybrid models on larger datasets to further validate and extend the proposed framework.

5.2. Model Interpretability and Clinical Reliability

Interpretability is a critical requirement for deploying deep learning models in clinical settings, as medical decisions must be transparent and explainable. In this study, interpretability is indirectly supported through the hierarchical feature learning capability of CNN architectures, where lower layers capture fundamental visual patterns such as edges and textures, and deeper layers extract clinically relevant features such as lesion asymmetry, border irregularity, and structural patterns. These characteristics are consistent with established dermatological criteria used in melanoma diagnosis. Furthermore, the reliability of the proposed framework is demonstrated through consistent performance across multiple evaluation metrics, including sensitivity, precision, and AUC, particularly for melanoma detection. The use of validation-based training, early stopping, and class weighting contributes to stable model behaviour and reduces the risk of overfitting. However, explicit explainability techniques such as Grad-CAM or SHAP were not implemented in this study, which limits the direct visualization of model decision regions [43]. Future work will incorporate explainable AI methods to provide visual interpretation of model predictions and enhance clinical trust. Additionally, evaluation on larger and more diverse datasets will further improve reliability and support real-world deployment of the proposed system.

5.3. Effect of Data Augmentation and Class Weighting

Data augmentation and class weighting were incorporated in this study to address the challenges of limited data availability and class imbalance in the MED-NODE dataset. These techniques contributed to improved model generalization and enhanced the detection of minority class melanoma samples, as reflected in the overall performance metrics. However, this study did not include dedicated ablation experiments to isolate the individual contributions of data augmentation and class weighting. Therefore, while their combined effect is evident, the exact quantitative impact of each component cannot be independently measured. Future work will include controlled ablation studies to systematically evaluate the effectiveness of these strategies and further optimize the training process.

6. Conclusion

This study proposed a deep learning model for classification of dermoscopic skin lesion using the MED-NODE dataset. The framework developed with transfer learning using pretrained CNN architectures such as VGG16, MobileNetV2 and EfficientNetB0. Fine-tuning, hierarchical feature extraction and silent feature preservation help the model to capture both low-level and high-level lesion characteristics. Among the models evaluated VGG16 achieved the highest accuracy, precision, F1-score and AUC. This suggests that deep learning models with pretrained features and selective fine-tuning are

suitable for small and imbalanced medical datasets. MobileNetV2 attains reasonable accuracy with low computational cost. cientNetB0 shows that increasing network size alone is insufficient to improve performance it also needs a suitable training process. The two major parts of training process such as data augmentation and class weighting help the model to avoid overfitting and perform better on unseen data. The results show that Model design, pre-processing and training strategies are as important as network architecture for achieving reliable automated skin lesion classification. A relatively small and imbalance dataset may restrict the broader applicability of these results. The frame work is expanded to a relatively large and more diverse dermoscopic datasets in future. It is also in consideration to evaluate hybrid or ensemble model and combine it with explainable AI techniques to make results more interpretable and relevant for clinical practice. Overall, the results of the model show that by carefully utilizing deep CNNs can effectively support automated melanoma detection that provide a strong foundation for the continues development of practical computer-aided dermatology tools.

Reference

1. Dlamini, Z., Hull, R., Molefi, T., Khanyile, R.E., Mbodi, L., Luvhengo, T.E., and Chikanza, I.C.: 'Understanding Melanoma: From Basics to Cutting-Edge Insights': 'Navigating Melanoma Treatment Challenges' (CRC Press, 2025), pp. 1-24
2. Zbrzezny, A.M., and Krzywicki, T.: 'Artificial intelligence in dermatology: A review of methods, clinical applications, and perspectives', Applied Sciences, 2025, 15, (14), pp. 7856
3. Abbas, T., Basit, M.K., Janjua, J.I., Naqvi, B.T., and Irfan, M.: 'Optimal Solution for Segmentation of Malignant Melanoma Dermoscopic Images', STATISTICS, COMPUTING AND INTERDISCIPLINARY RESEARCH, 2023, 5, (2), pp. 39-49
4. Matsuzaka, Y., and Yashiro, R.: 'The diagnostic classification of the pathological image using computer vision', Algorithms, 2025, 18, (2), pp. 96
5. Ahmad, Z., Rahim, S., Zubair, M., and Abdul-Ghafar, J.: 'Artificial intelligence (AI) in medicine, current applications and future role with special emphasis on its potential and promise in pathology: present and future impact, obstacles including costs and acceptance among pathologists, practical and philosophical considerations. A comprehensive review', Diagnostic pathology, 2021, 16, (1), pp. 24
6. Noronha, S.S., Mehta, M.A., Garg, D., Kotecha, K., and Abraham, A.: 'Deep learning-based dermatological condition detection: A systematic review with recent methods, datasets, challenges, and future directions', IEEE Access, 2023, 11, pp. 140348-140381
7. Jatale, K.A., and Deshpande, D.S.: 'Advances in Skin Disease Detection and Classification Using Convolutional Neural Networks: A Comprehensive Literature Survey', in Editor (Ed.)^(Eds.): 'Book Advances in Skin Disease Detection and Classification Using Convolutional Neural Networks: A Comprehensive Literature Survey' (IEEE, 2025, edn.), pp. 1-7
8. Ihnaini, B., Khan, M.A., Khan, T.A., Abbas, S., Daoud, M.S., Ahmad, M., and Khan, M.A.: 'A smart healthcare recommendation system for multidisciplinary diabetes patients with data fusion based on deep ensemble learning', Computational Intelligence and Neuroscience, 2021, 2021, (1), pp. 4243700
9. Abbas, T., Linjawi, M., Iqbal, I., Alghushairy, O., Alsini, R., and Daud, A.: 'Significance of unifying semi and self-supervision for the radical improvement of medical imaging: a collaborative research effort', Biomedical Signal Processing and Control, 2026, 111, pp. 108391
10. Zaidi, A.R., Abbas, T., Daud, A., Alghushairy, O., Dawood, H., and Sarwar, N.: 'Enhancing Android malware detection with XGBoost and convolutional neural networks', Computers, Materials & Continua, 2025, 84, (2), pp. 3281-3304
11. Abdikenov, B., Zhaksylyk, T., Shortanbaiuly, O., Orazayev, Y., Makhanov, N., Karibekov, T., Suvorov, V., Imasheva, A., Zhumagozhayev, K., and Seitova, A.: 'Future of breast cancer diagnosis: A review of DL and ML applications and emerging trends for multimodal data', IEEE Access, 2025
12. Mienye, I.D., Swart, T.G., Obaido, G., Jordan, M., and Ilono, P.: 'Deep convolutional neural networks in medical image analysis: A review', Information, 2025, 16, (3), pp. 195
13. Salehi, A.W., Khan, S., Gupta, G., Alabduallah, B.I., Almjally, A., Alsolai, H., Siddiqui, T., and Mellit, A.: 'A study of CNN and transfer learning in medical imaging: Advantages, challenges, future scope', Sustainability, 2023, 15, (7), pp. 5930
14. Jain, U., Chouhan, S.P., Chakraborty, R., and Bhattacharya, M.: 'Ensemble-Based Deep Learning Framework for Multi-Class Skin Lesion Diagnosis with Class Imbalance Mitigation', in Editor (Ed.)^(Eds.): 'Book Ensemble-Based Deep Learning Framework for Multi-Class Skin Lesion Diagnosis with Class Imbalance Mitigation' (IEEE, 2025, edn.), pp. 6838-6845
15. Naseri, H., and Safaei, A.A.: 'Diagnosis and prognosis of melanoma from dermoscopy images using machine learning and deep learning: a systematic literature review', BMC cancer, 2025, 25, (1), pp. 75
16. Maharana, K., Mondal, S., and Nemade, B.: 'A review: Data pre-processing and data augmentation techniques', Global Transitions Proceedings, 2022, 3, (1), pp. 91-99
17. Akgündoğdu, A., and Çelikbaş, Ş.: 'Explainable deep learning framework for brain tumor detection: Integrating LIME, Grad-CAM, and SHAP for enhanced accuracy', Medical Engineering & Physics, 2025, pp. 104405
18. Ruga, T., Zumpano, E., Vocaturo, E., Caroprese, L., and Arlia, C.: 'Bias in dermatological datasets: A critical analysis of the underrepresentation of dark skin tones in melanoma classification images', in Editor (Ed.)^(Eds.): 'Book Bias in dermatological datasets: A critical analysis of the underrepresentation of dark skin tones in melanoma classification images' (Springer, 2025, edn.), pp. 434-448

19. Arik, K.: 'From CNNs to Vision Transformers: A Survey on Addressing Domain Shift and Adaptation Concepts in Medical Image Analysis', Archives of Computational Methods in Engineering, 2026, pp. 1-40
20. Islam, S., Wishart, G.C., Walls, J., Hall, P., G. Seco de Herrera, A., Gan, J.Q., and Raza, H.: 'Advancing skin cancer detection through deep learning and fusion of patient metadata and skin lesion images', Scientific Reports, 2026
21. Priyadarsini, R.N., Tyagi, B., and Priyadharsini, M.: 'An attention based optimized network for the classification of skin lesions', Scientific Reports, 2026
22. Vieira, J., Mendonça, F., and Morgado-Dias, F.: 'Deep Learning Approaches for Skin Lesion Detection', Electronics, 2025, 14, (14), pp. 2785
23. Chiu, T.-M., Chi, I.-C., Li, Y.-C., and Tseng, M.-H.: 'Deep ensemble learning for multiclass skin lesion classification', Bioengineering, 2025, 12, (9), pp. 934
24. Zhang, X., Liu, Y., Ouyang, G., Chen, W., Xu, A., Hara, T., Zhou, X., and Wu, D.: 'DermViT: Diagnosis-guided vision transformer for robust and efficient skin lesion classification', Bioengineering, 2025, 12, (4), pp. 421
25. Aksoy, S., Demircioglu, P., and Bogrekci, I.: 'Deep learning-based web application for automated skin lesion classification and analysis', Dermato, 2025, 5, (2), pp. 7
26. Erbay, H., Abulgasim, Y.M., Özer, D.E., and Ertürk, F.: 'Enhancing multi-class skin lesion diagnosis through ensemble learning of CNN and transformer architectures', Engineering Science and Technology, an International Journal, 2025, 70, pp. 102145
27. Wang, D., Liu, Q., Hu, Y., and Liu, H.: 'Multi-Scale Attention Fusion with Lesion-Area Focus for Knowledge-Enhanced Dermoscopic Skin Lesion Classification', Applied Sciences, 2025, 15, (24), pp. 12952
28. Junejo, R.H., Abbas, Q., Awais, M., Akram, T., and Aldajani, M.B.: 'Federated ViT: A distributed deep learning framework for skin cancer classification', IEEE Access, 2025
29. Aldeen, Z.D., and Ibrahim, L.M.: 'Emotion Recognition in Customer Service Systems: A Human-Centered Software Engineering Approach to Enhancing User Experience', in Editor (Ed.)^(Eds.): 'Book Emotion Recognition in Customer Service Systems: A Human-Centered Software Engineering Approach to Enhancing User Experience' (Springer, 2025, edn.), pp. 298-311
30. Nawaz, K., Zanib, A., Shabir, I., Li, J., Wang, Y., Mahmood, T., and Rehman, A.: 'Skin cancer detection using dermoscopic images with convolutional neural network', Scientific Reports, 2025, 15, (1), pp. 7252
31. Rahat, M., Islam, M.M.U., Akon, M.Z., and Faisal, D.R.H.: 'A Deep Feature Extraction and Optimization Framework for ECG Classification via Pretrained Vision Networks', in Editor (Ed.)^(Eds.): 'Book A Deep Feature Extraction and Optimization Framework for ECG Classification via Pretrained Vision Networks' (Springer, 2025, edn.), pp. 374-389
32. Fushshilat, I., Rizal, A., and Cahyadi, W.A.: 'An optimized deep learning pipeline for used IC classification integrating DenseNet121 with fine-tuning', in Editor (Ed.)^(Eds.): 'Book An optimized deep learning pipeline for used IC classification integrating DenseNet121 with fine-tuning' (IEEE, 2025, edn.), pp. 311-316
33. Elangbam, B., and Mandala, S.: 'Image Processing and Machine Learning Approaches for Skin Diseases Diagnosis: Trends, Challenges, and Future Directions', in Editor (Ed.)^(Eds.): 'Book Image Processing and Machine Learning Approaches for Skin Diseases Diagnosis: Trends, Challenges, and Future Directions' (IEEE, 2025, edn.), pp. 669-675
34. Alirr, O.I.: 'A Large-Kernel and Scale-Aware 2D CNN with Boundary Refinement for Multimodal Ischemic Stroke Lesion Segmentation', Eng, 2026, 7, (2), pp. 59
35. Khan, M.A., Mazhar, T., Ali, M.D., Khattak, U.F., Shahzad, T., Saeed, M.M., and Hamam, H.: 'Automatic melanoma and non-melanoma skin cancer diagnosis using advanced adaptive fine-tuned convolution neural networks', Discover Oncology, 2025, 16, (1), pp. 645
36. Fatima, S., Akram, M.U., Mohammad, S., and Ahmed, S.B.: 'Deep learning in dermatopathology: applications for skin disease diagnosis and classification', Discover Applied Sciences, 2025, 7, (9), pp. 1006
37. Alrabai, A., Echtioui, A., and Kallel, F.: 'Exploring pre-trained models for skin cancer classification', Applied System Innovation, 2025, 8, (2), pp. 35
38. Tran, A.T., Zeevi, T., and Payabvash, S.: 'Strategies to improve the robustness and generalizability of deep learning segmentation and classification in neuroimaging', BioMedInformatics, 2025, 5, (2), pp. 20
39. Borrohou, S., Fissoune, R., and Badir, H.: 'Critical role of data transformation in preprocessing: methods, algorithms, and challenges', in Editor (Ed.)^(Eds.): 'Book Critical role of data transformation in preprocessing: methods, algorithms, and challenges' (Springer, 2024, edn.), pp. 108-122

40. Sharma, A.K., Bhanodia, P.K., and Bhushan, M.: 'A hybrid deep learning approach for skin lesion classification using convolutional neural networks and transformers', Network Modeling Analysis in Health Informatics and Bioinformatics, 2026, 15, (1), pp. 119
41. Jabber, A.A., Shadeed, G.A., Salim, N., and Dibs, H.: 'Automated skin lesion diagnosis and classification using k-mean, lab-color-space segmentation and deep learning', in Editor (Ed.)^(Eds.): 'Book Automated skin lesion diagnosis and classification using k-mean, lab-color-space segmentation and deep learning' (Springer, 2025, edn.), pp. 43
42. Ye, C., Li, J., and Shuai, Q.: 'Evaluating the Performance and Clinical Applications of Multiclass Deep Learning Models for Skin Cancer Pathology Diagnosis (ISIC): A Comparative Analysis of CNN, ViT, and VLM', in Editor (Ed.)^(Eds.): 'Book Evaluating the Performance and Clinical Applications of Multiclass Deep Learning Models for Skin Cancer Pathology Diagnosis (ISIC): A Comparative Analysis of CNN, ViT, and VLM' (2025, edn.), pp. 92-103
43. Kohan, A., Zahedi, A., Alizadehsani, R., Tan, R.S., and Acharya, U.R.: 'Application of explainable artificial intelligence (XAI) techniques in patients with intracranial hemorrhage: A systematic review', Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2025, 15, (3), pp. e70031