Comparative Analysis of Transfer Learning Models for Multi-Class Skin Disease Classification on HAM10000

Partha Jyoti Roy[1,*] and Faisal Islam[2]

[1,2] Department of Electrical and Electronic Engineering, Khulna University of Engineering & Technology (KUET), Khulna 9203, Bangladesh
roypartha966@gmail.com[1,*], faisalislam620@gmail.com[2]

Abstract. Skin diseases are a major public health issue worldwide, requiring early diagnosis and intervention to prevent serious consequences, including melanoma-related deaths. However, the traditional dermatoscopic diagnosis of skin diseases is often time-consuming, subjective, and depends on the availability of skilled dermatologists. To improve the diagnosis of skin diseases, the current research provides a comprehensive investigation of the latest transfer learning-based convolutional neural network (CNN) models for multi-class skin disease classification using dermatoscopic images. Six pre-trained deep learning models, including InceptionResNetV2, NASNetLarge, EfficientNetV2L, DenseNet201, ResNet101V2, and VGG16, were fine-tuned and tested on the HAM10000 dataset, consisting of 10,015 images divided into seven classes of skin disease. A series of experiments were conducted to evaluate the effectiveness of the proposed models in multi-class skin disease classification. Experimental results show that fine-tuning significantly improves model convergence and generalization. Of all the models tested, EfficientNetV2L achieved the highest classification accuracy, test accuracy of 87.92%, lowest loss of 4.76%, and highest precision, recall, and F1-score. The experimental results support the fact that the latest CNN models, which are optimized for scaling and feature extraction, are better than traditional models for multi-class skin disease classification. This research provides valuable insights for choosing effective deep learning models for multi-class skin disease classification, and the results show the effectiveness of EfficientNet-based models for developing reliable multi-class skin disease classification models.

Keywords: Transfer learning, Fine-tuning, Skin disease, InceptionResNetV2, NASNetLarge, EfficientNetV2L, DenseNet201, ResNet101V2, VGG16.

1 Introduction

Skin diseases are a significant burden on the health of the world's population, with millions of people diagnosed annually, ranging from benign diseases to life-threatening cancer, such as melanoma. Timely and accurate diagnosis is critical in order to avoid the progression of the disease and improve patient outcomes [1]. However, traditional

M. S. Arefin et al. (eds.), *Proceedings of the International Conference on Intelligent Data Analysis and Applications (IDAA 2025)*, Advances in Intelligent Systems Research 206,
https://doi.org/10.2991/978-94-6239-664-7_97

diagnosis methods, such as dermoscopic examination, are often time-consuming, subjective, and highly dependent on skilled dermatologists, who are not always readily available in rural and developing regions.

Automated skin disease detection has been recognized as a potential solution to conventional diagnostic procedures with recent advances in deep learning technology and computer vision. Convolutional neural networks (CNNs) have demonstrated remarkable performance in dermoscopic image analysis with their strong ability to learn complex visual patterns, such as texture, color distribution, and structural abnormalities associated with various skin conditions. The feature learning ability of CNNs has made them outperform conventional machine learning approaches.

Transfer learning plays a vital role in the automatic classification of skin diseases, especially in medical image processing scenarios where annotated image data is limited in quantity. Using pre-trained convolutional neural networks and fine-tuning these models on smaller datasets like HAM10000 can significantly reduce training time and improve the accuracy of the models while avoiding overfitting. In this study, the HAM10000 dataset consisting of 10,015 images of seven different types of skin lesion images is used for evaluating the efficiency of transfer learning in automatic classification of skin diseases [2].

The remainder of this paper is organized as follows: Section 2 discusses the related work on the automated detection of skin disease, while Section 3 discusses the proposed methodology, including the dataset, preprocessing steps, data augmentation methods, transfer learning models, and fine-tuning methods. In addition, Section 4 discusses the analysis of the results and performance evaluation, while Section 5 provides concluding remarks and an overview of the major contributions of the proposed approach, as well as future research directions.

2 Related Works

Skin disease classification has been recognized as a key issue in medical image analysis, primarily for the purpose of aiding dermatologists in the early and precise diagnosis of skin cancers such as melanoma. In recent times, a large number of deep learning-based techniques have been developed to increase the reliability of computer-aided diagnostic systems.

In reference [3], the authors propose a deep ensemble framework in skin disease classification by employing three pre-trained convolutional neural networks, namely VGG16, InceptionV3, and ResNet50, based on their evaluation metrics on the ISIC dataset. The ensemble of these models achieved an accuracy of 97% in classifying the balanced dataset, significantly higher than each of their individual results. This study proves that an ensemble of different architectures of CNN can utilize their capabilities of feature extraction in an efficient way.

Another significant contribution is reported in [4], where a transfer learning-based Xception model is proposed, and attention mechanisms are incorporated into it. The authors have used self-attention, hard-attention, and soft-attention mechanisms and compared their performance on the HAM10000 dataset, showing that the self-attentionenhanced Xception model has the highest accuracy of 94.11%, while the baseline model has an accuracy of 91.05%. In addition, a comparative analysis of the proposed models, namely, DenseNet201, VGG16, Inception-ResNetV2, and ResNet50, has been performed, showing that attention-based models have higher classification accuracy and recall, especially when classifying benign and malignant lesions.

The research of Rashid et al. [5] was carried out on the topic of transfer learning for melanoma detection using the ISIC-2020 dataset. A suite of deep convolutional neural networks (CNN), including Res-Net152V2, DenseNet201, InceptionV3, and NASNetLarge, was used for this purpose. In order to reduce the problem of class imbalance in the data, a variety of data augmentation strategies were applied, showing improved results in terms of generalization. Out of all the networks used in this research, the highest accuracy was obtained using the NASNetLarge network due to its neural architecture search optimization. In addition, the Res-Net152V2 and DenseNet201 networks also provided good results due to the efficient flow of gradients in the networks. The research concludes that deeper networks are highly effective for fine-grained classification tasks in medical images.

Although initial research works have mainly focused on mainstream architectures such as VGG16, ResNet, and Inception, recent research works show that sophisticated architectures such as Xception, DenseNet201, and NASNetLarge perform better, especially in combination with ensemble learning and attention. These architectures can well describe the subtle changes in dermoscopic images and thus provide more accurate and robust diagnoses. However, there is still less research in comparing these state-of-the-art architectures under controlled experimental settings. The relationship between transfer learning, complexity, and performance is still unclear.

This study seeks to conduct an extensive evaluation of some of the current models of convolutional neural networks (CNNs) like Inception, ResNetV2, ResNet152V2, Xception, DenseNet201, VGG16, and NASNetLarge to identify the most reliable model for the automatic classification of skin diseases.

3 Methodology

3.1 Image Classification

This research employs a supervised image classification approach to automatically detect and classify skin diseases from dermoscopic images. In this approach, a single label is assigned to an image based on features learned from a dataset of images. In this way, a classification approach is used to map dermoscopic images into meaningful disease classes.

Image classification plays an important role in the early detection of various skin diseases, allowing doctors to make decisions based on predictions that are precise and reproducible. Image classification is a significant problem in computer vision and deep learning. This problem has been successfully used in many fields, including medical diagnoses, object recognition, and disease detection. In dermatology, deep CNNs have shown their power in dealing with various problems, particularly in their ability to learn hierarchical representations of textures, colors, and structural patterns of different diseases.

3.2 Image Dataset

The experiments conducted in this study used the HAM10000 dataset, which consists of 10,015 dermoscopic images of seven classes of skin diseases: actinic keratoses, melanocytic nevi, benign keratosis-like lesions, melanoma, basal cell carcinoma, dermatofibroma, and vascular lesions. The dataset contains images of both benign and malignant conditions, thus allowing multi-class classification of skin diseases.

The dataset has a significant class imbalance. The class of melanocytic nevi has the highest number of instances, contributing 66.9% of the dataset or 6,705 images. Dermatofibroma has the smallest number of instances, with only 115 images or 1.1% of the dataset. The class of vascular lesions has a small contribution to the dataset, contributing only 1.4% of the dataset. This is a problem that requires the application of specific techniques to avoid bias towards the major classes.

The images used in the dataset are derived from diverse clinical resources, which makes them heterogeneous with regard to resolution, illumination conditions, skin tones, and backgrounds. In order to make the images compatible with a convolutional neural network architecture, all images used in this study were resized to 224 × 224 pixels to ensure that all images have the same resolution. The images were saved in RGB format to ensure that critical color and texture information were retained.

In order to handle the issue of class imbalance and to increase the generality of the model, the training process involves class weighting, focal loss, and various data augmentation strategies. These increase the robustness of the model to changes in lighting conditions, skin tones, and acquisition settings, thus making the proposed classification system reliable in a real-world clinical setting.

3.3 Dataset Preprocessing and Data Augmentation

For dataset preprocessing, the given CSV file was used to map the dermoscopic images with their respective class labels. For this purpose, the dataset was split into training, validation, and test subsets in the proportion of 80%, 10%, and 10%, respectively, by using stratified sampling.

All the images were resized to a fixed size of 224x224 pixels for the input of the different architectures of the Convolutional Neural Network. Normalization of the pixel values was performed by scaling the values in the range of 0 and 1.

In order to improve the robustness of the models and avoid overfitting, data augmentation was performed extensively only on the training data. The data augmentation methods included random flipping, rotations up to 20°, zoom up to 20%, and scaling by a factor of 1/255. These data augmentation methods improve the variety of the data and the ability of the model to generalize the data. However, for the validation and test data, scaling by a factor of 1/255 was performed.

This standardized preprocessing and augmentation pipeline effectively addresses class imbalance, improves generalization, and ensures reliable and reproducible classification results.

3.4 Transfer Learning Methods Applied to Image Classification

Transfer learning is a technique that employs pre-trained knowledge from models trained on large datasets to improve performance on specific domain tasks with limited labeled data. In medical image analysis, this method is particularly effective as it reduces training time while enhancing feature extraction and classification accuracy. In this study, various state-of-the-art CNN architectures were pre-trained for skin disease classification.

InceptionResNetV2, introduced by Google in 2016, combines the multi-scale feature extraction capability of Inception modules with residual connections from ResNet architectures. This hybrid design mitigates vanishing gradient issues and enables faster convergence in very deep networks. With over 150 layers and approximately 55 million parameters, InceptionResNetV2 has demonstrated strong performance in fine-grained classification and medical imaging applications.

NASNetLarge, proposed by Google in 2017, applies reinforcement learning to Neural Architecture Search (NAS) to learn the most efficient architecture for the network. NASNetLarge is composed of modular normal and reduction cells, consisting of over 88 million parameters and achieving high accuracy on large datasets. Its scalable and efficient architecture makes NASNetLarge most effective for fine-grained image classification problems [6].

VGG16 has been widely known for its simplicity and efficiency since its introduction in 2014 by the Visual Geometry Group at Oxford University. The network consists of 16 trainable layers that use small 3x3 convolutional kernels and 2x2 max pooling layers. Despite having a massive number of parameters, approximately 138 million, VGG16 remains a powerful baseline for image classification and feature extraction tasks in medical imaging scenarios [7].

ResNet101V2, proposed by Microsoft Research in 2016, is an extension of the original ResNet architecture, where pre-activation residual blocks are used, which means that batch normalization and ReLU activation are performed before the convolutional layers. This has been found to improve stability during the training of deep networks. ResNet101V2 has 101 layers and has been used for a variety of applications, including medical imaging, object detection, and fine-grained classification [8].

DenseNet201 was proposed by researchers from Cornell University and Facebook AI Research in 2017, and the architecture is based on dense connectivity, meaning that each layer is connected to all the previous layers. This design allows for the efficient reuse of features and gradients, which is why the network can be learned with just 20 million parameters. DenseNet201 has been shown to be effective for image and medical image classification [9].

EfficientNetV2L, proposed by Google in 2021, has been designed to achieve high accuracy with faster training and better computational efficiency. The large version uses compound scaling methods to adjust the depth, width, and resolution of the input image. It also uses fused MBConv, progressive learning, and better regularization. It has over 120 million parameters and achieves state-of-the-art accuracy on large data sets, making it a better option for medical image classification and fine-grained recognition tasks [10].

3.5 Fine Tuning

In other words, fine-tuning is a transfer learning technique in which a pre-trained deep network is adapted to a specific, novel domain for a specific task by retraining a subset or all of its parameters on labeled data for the target domain. In convolutional neural networks, the lower layers tend to learn generic visual features such as edges, textures, and basic shapes, while the higher layers learn features specific to the target classification task.

For the fine-tuning of the model, as implemented in the current study, the early convolutional layers are first frozen, aiming at retaining the general features learned during the pretraining phase. Then, the subsequent layers, as well as the newly added classification layers, are unfrozen and retrained, allowing the model to learn and adapt to the characteristics of the dermoscopic images, while avoiding the loss of the significant general knowledge acquired during the initial phase.

In order to guarantee stable convergence and prevent catastrophic forgetting, lower learning rates are used during fine-tuning. This optimized optimization process allows for gradual adaptation, thus improving discrimination between visually similar classes of skin lesions. The fine-tuning process is particularly useful for medium-sized medical image datasets like HAM10000 because it improves classification performance, reduces overfitting, and improves generalization.

4 Result and Discussion

4.1 Research Design

The primary objective of this research is to develop and evaluate an automated system for dermatoscopic skin disease classification using modern transfer learning approaches. For this purpose, six different pre-trained convolutional neural network architectures, namely InceptionResNetV2, NASNetLarge, EfficientNetV2L, DenseNet201, ResNet101V2, and VGG16, have been comprehensively tested using the HAM10000 dataset.

To evaluate the performance of each of these architectures, conventional evaluation parameters such as accuracy, precision, recall, F1-score, and loss have been used. In addition to these parameters, ROC curves have also been employed to determine the discriminative power of each of these architectures for all disease classes. At the same time, confusion matrices have also been used to analyze the results obtained using each of these architectures for skin disease classification.

In the initial stage, all models were pre-trained on ImageNet weights, which were obtained from the Keras/TensorFlow framework. The weights for the convolutional base of all models were frozen, and only the weights for the classification head were trained. Global Average Pooling (GAP) was applied to the convolutional base for all models, and a fully connected layer with Softmax activation function was added for multi-class classification on the seven-class HAM10000 dataset.

Hyperparameter tuning was done to optimize the model's performance. The learning rate was set to 0.0001 to control weight updates, and the batch size was set to 16. The number of epochs was set to 20 to avoid underfitting and overfitting. To deal with the extreme class imbalance, class weights were used. Focal loss was used as the objective function. The Adam optimizer was used because of its adaptive learning rate. The accuracy and loss were constantly monitored during training and validation.

Yet another refinement was incorporated in the form of a fine-tuning phase, where further adaptation of the pre-trained models was achieved in consideration of the dermatoscopic image characteristics. For this purpose, the previously frozen layers of the convolutional layers were unfrozen, allowing all parameters of the models to be learned during training with the aid of the HAM10000 dataset. For this purpose, the learning rate was decreased to 0.00001 in order to avoid catastrophic forgetting and overfitting. The training was continued for an additional 20 epochs and was coupled with the ReduceLROnPlateau learning rate scheduler.

Table 1. Hyperparameter values before and after fine-tuning.

Hyperparameter	Before fine-tuning	After fine-tuning
Learning Rate	0.0001	0.00001
Batch Size	16	16
Epochs	20	20
Optimizer	Adam	Adam
Loss Function	Focal Loss + Class Weights	Focal Loss + Class Weights
Activation (Last Layer)	Softmax	Softmax
Layer Freezing	Base Frozen, Head Trainable	All Layers Trainable
LR Scheduler	None	ReduceLROnPlateau

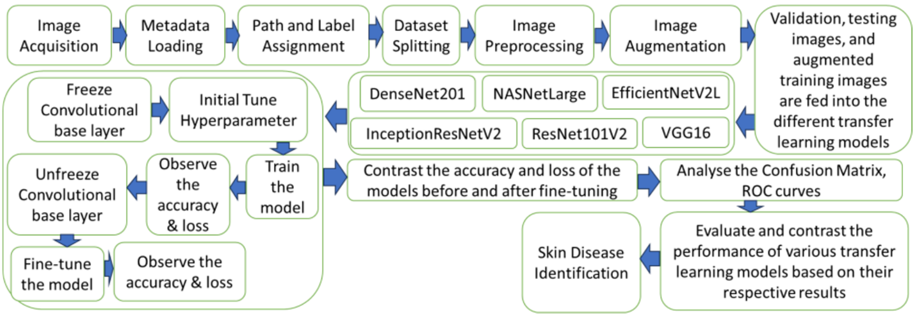


Fig. 1. An illustration of the entire research work.

The complete set of hyperparameters used before and after fine-tuning is summarized in **Table 1**. The overall research workflow is illustrated in **Fig. 1**, depicting the sequential process of data acquisition, preprocessing, augmentation, transfer learning, fine-tuning, and performance evaluation. Final model assessment includes ROC curves, confusion matrices, and comprehensive performance metrics, ensuring a systematic and reproducible approach to achieving optimal multi-class skin disease classification.

4.2 Experimental Results

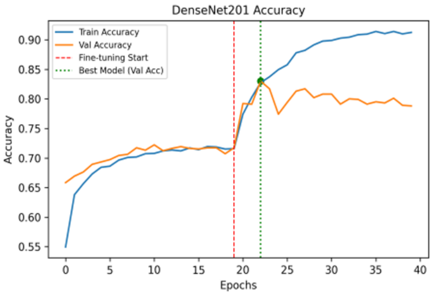
Fig. 2 illustrates the training and validation accuracy curves of the six transfer learning models before and after fine-tuning. Fine-tuning improved the classification accuracy of all tested architectures: InceptionResNetV2, NASNetLarge, EfficientNetV2L, DenseNet201, ResNet101V2, and VGG16. However, the degree of improvement and training stability differed among the models.

EfficientNetV2L and NASNetLarge showed the highest training stability and consistency, with smooth accuracy curves and small oscillations. ResNet101V2 also showed stable convergence with small oscillations, while VGG16 had moderate oscillations during training. However, DenseNet201 had large oscillations and signs of overfitting in the latter stages of training. After fine-tuning, all models showed improved convergence and higher validation accuracy, confirming that fine-tuning improved high-level feature extraction. Among the models, EfficientNetV2L showed the best trade-off between training stability and generalization, while DenseNet201 greatly benefited from fine-tuning adjustments.

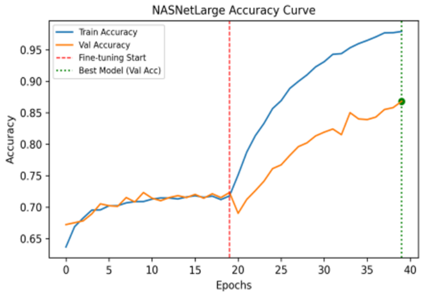
The training and validation loss curves for the models are shown in Fig. 3. After fine-tuning, there was a decrease in both training and validation loss for all models. EfficientNetV2L, NASNetLarge, and ResNet101V2 showed smooth loss curves, indicating stable optimization. VGG16 showed small oscillations, while DenseNet201 showed a slight increase in validation loss in the latter stages of training, indicating mild overfitting beyond its optimal point. There were small loss spikes at the transition points during fine-tuning of NASNetLarge, indicating sensitivity of the architecture to unfreezing model parameters.

Before fine-tuning, all models were trained for 20 epochs with the same hyperparameters to compare them equally. The baseline training and validation performance of all models is presented in Table 2 to compare the effect of fine-tuning. After fine-tuning, the models were trained for an additional 20 epochs, and the improvements in accuracy and loss are presented in Table 3. For all models, fine-tuning resulted in higher training and validation accuracy and lower loss, confirming its effectiveness in improving adaptability to dermoscopic image data.

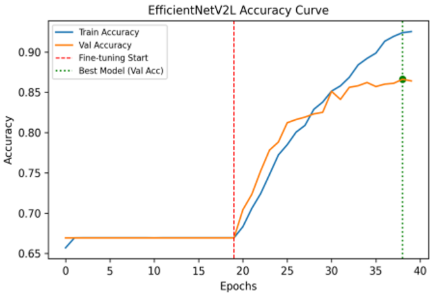
Table 4 shows the performance of the six transfer learning models on the test set. EfficientNetV2L was at the top, achieving the best test accuracy of 87.92% and the lowest test loss of 4.76%. Second on the list was NASNetLarge, followed by ResNet101V2, which achieved 84.93% accuracy and 6.09% test loss, respectively. Then came VGG16, which gave an average performance, followed by DenseNet201 and InceptionResNetV2, which showed lower accuracy and higher test loss, indicating that these models were not reliable for clinical prediction purposes. Therefore, it can be inferred that modern architectures, which have been refined for feature extraction and scaling, perform better than traditional CNN architectures for multi-class skin disease classification.



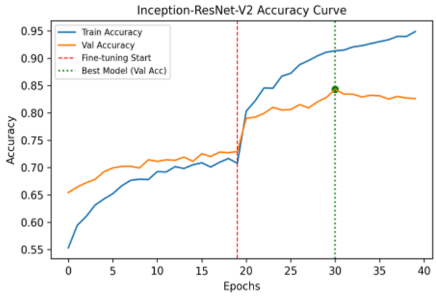
(a) DenseNet201



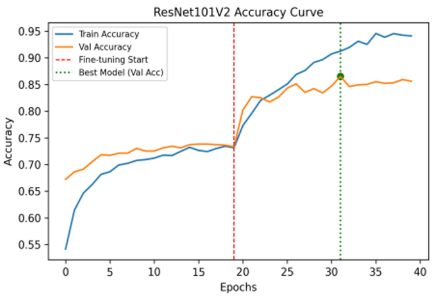
(b) NASNetLarge



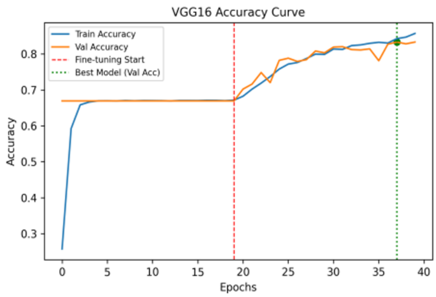
(c) EfficientNetV2L



(d) InceptionResNetV2



(e) ResNet101V2

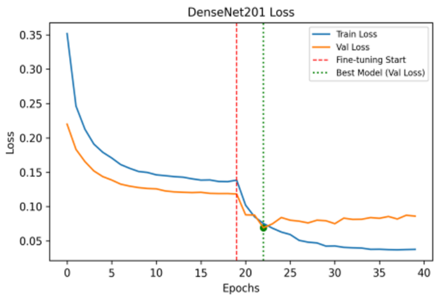


(f) VGG16

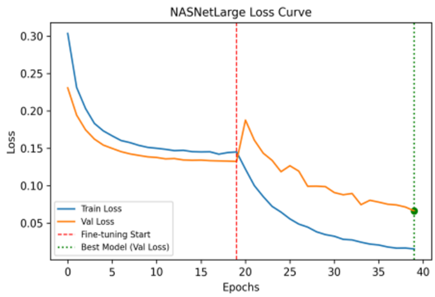
Fig. 2. Training and validation accuracy curves of the six transfer learning models before and after fine-tuning on the HAM10000 dataset.

Additionally, we considered precision, recall, and F1-score, which are shown in Table 4. EfficientNetV2L scored the highest in all three, indicating that it has strong class-wise discrimination and balanced performance. Then came NASNetLarge and ResNet101V2, while InceptionResNetV2 showed the lowest precision, and DenseNet201 showed lower recall and F1-score, particularly for minority classes.

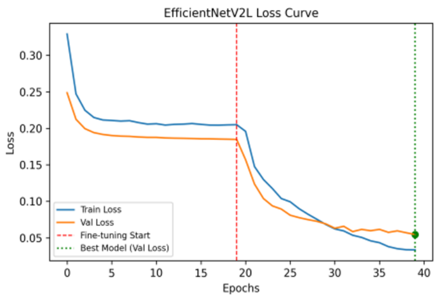
Fig. 4 presents the confusion matrices for all six models, which provide an in-depth analysis of the performance of the models. EfficientNetV2L, NASNetLarge, and ResNet101V2 showed higher correct classifications for common and rare diseases, while InceptionResNetV2 and DenseNet201 showed reasonable performance for majority classes, while they were likely to classify less represented classes, i.e., actinic keratoses were classified as benign keratosis-like lesions or melanoma.



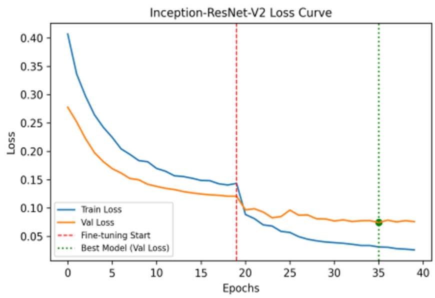
(a) DenseNet201



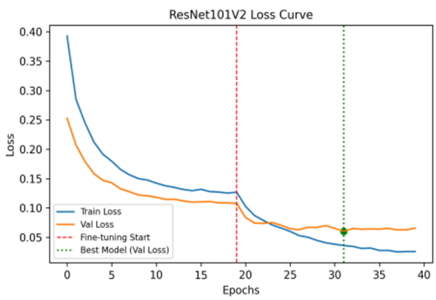
(b) NASNetLarge



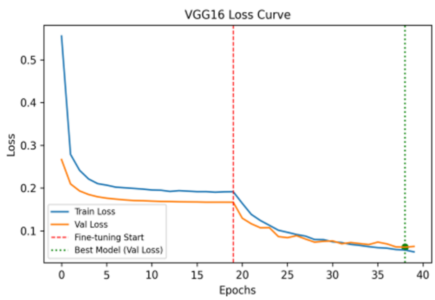
(c) EfficientNetV2L



(d) InceptionResNetV2



(e) ResNet101V2



(f) VGG16

Fig. 3. Training and validation loss curves of the six transfer learning models before and after fine-tuning on the HAM10000 dataset.

Fig. 5 presents the ROC curves for all six models, along with the Micro-AUC and Macro-AUC scores. All models showed high AUC values, indicating strong class-wise discrimination. Moreover, the Macro-AUC values showed strong performance for majority and minority classes, while the Micro-AUC values showed strong overall performance for all classes.

Additionally, we considered representative test images for all classes, shown in Fig. 6, and the performance is summarized in Table 5. EfficientNetV2L showed perfect performance for all seven representative images, while InceptionResNetV2 showed incorrect classifications for basal cell carcinoma and melanoma, which were classified as benign keratosis-like lesions, while DenseNet201 showed incorrect classifications for basal cell carcinoma and dermatofibroma.

VGG16 had the highest misclassification rate, which mostly confused melanoma with melanocytic nevi and misclassified vascular as well as keratosis-like features. NASNetLarge and ResNet101V2 had a few misclassifications, which were mostly of melanoma and basal cell carcinoma.

Table 2. Training and validation accuracy and loss of all transfer learning models before fine-tuning on the HAM10000 dataset.

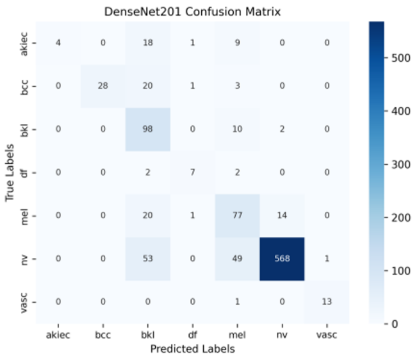
Models	Training Accuracy	Training Loss	Validation Accuracy	Validation Loss
DenseNet201	69.20	16.88	16.88	13.64
NASNetLarge	70.31	16.74	16.74	14.90
VGG16	64.49	22.18	22.18	17.86
ResNet101V2	69.51	17.21	17.21	13.54
InceptionResNetV2	67.09	20.30	20.30	15.90
EfficientNetV2L	66.89	21.62	21.62	19.25

Table 3. Training and validation accuracy and loss of all transfer learning models after fine-tuning.

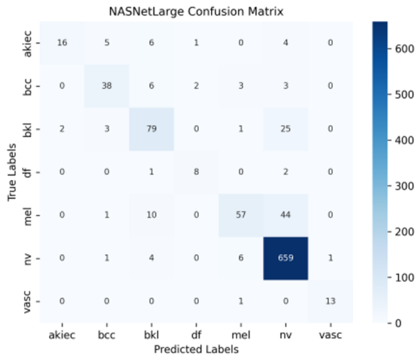
Models	Training Accuracy	Training Loss	Validation Accuracy	Validation Loss
DenseNet201	87.93	5.16	79.99	8.14
NASNetLarge	90.69	4.46	80.03	10.40
VGG16	79.23	8.49	79.06	8.16
ResNet101V2	88.75	4.79	84.19	6.74
InceptionResNetV2	89.73	4.62	81.94	8.29
EfficientNetV2L	83.15	7.82	82.10	7.71

Table 4. Test performance comparison of all evaluated models, including accuracy, loss, precision, recall, and F1-score.

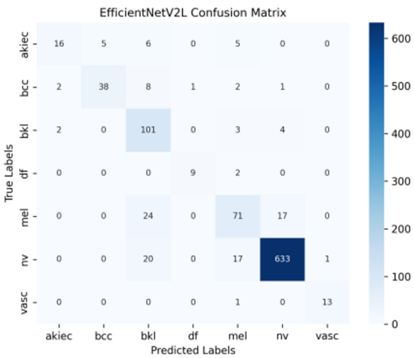
Models	Test Accuracy	Test Loss	Precision	Recall	F1 Score
DenseNet201	79.34	8.14	0.8638	0.7934	0.8024
NASNetLarge	86.83	7.39	0.8647	0.8683	0.8588
VGG16	84.43	5.71	0.8472	0.8443	0.8389
ResNet101V2	84.93	6.09	0.8571	0.8493	0.8506
InceptionResNetV2	83.03	8.03	0.8467	0.8303	0.8291
EfficientNetV2L	87.92	4.76	0.8897	0.8792	0.8799



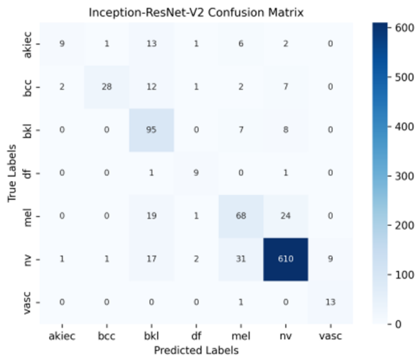
(a) DenseNet201



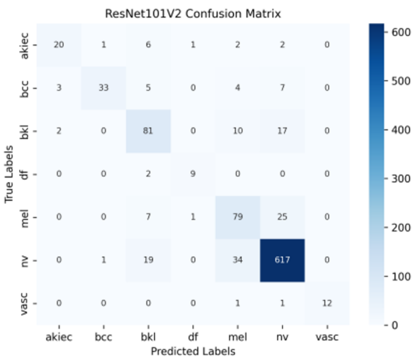
(b) NASNetLarge



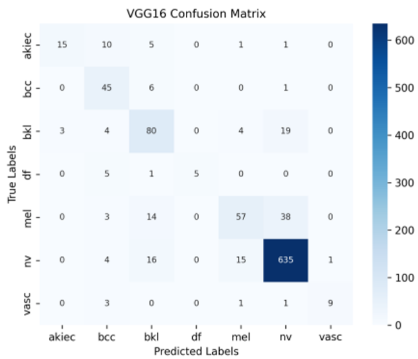
(c) EfficientNetV2L



(d) InceptionResNetV2

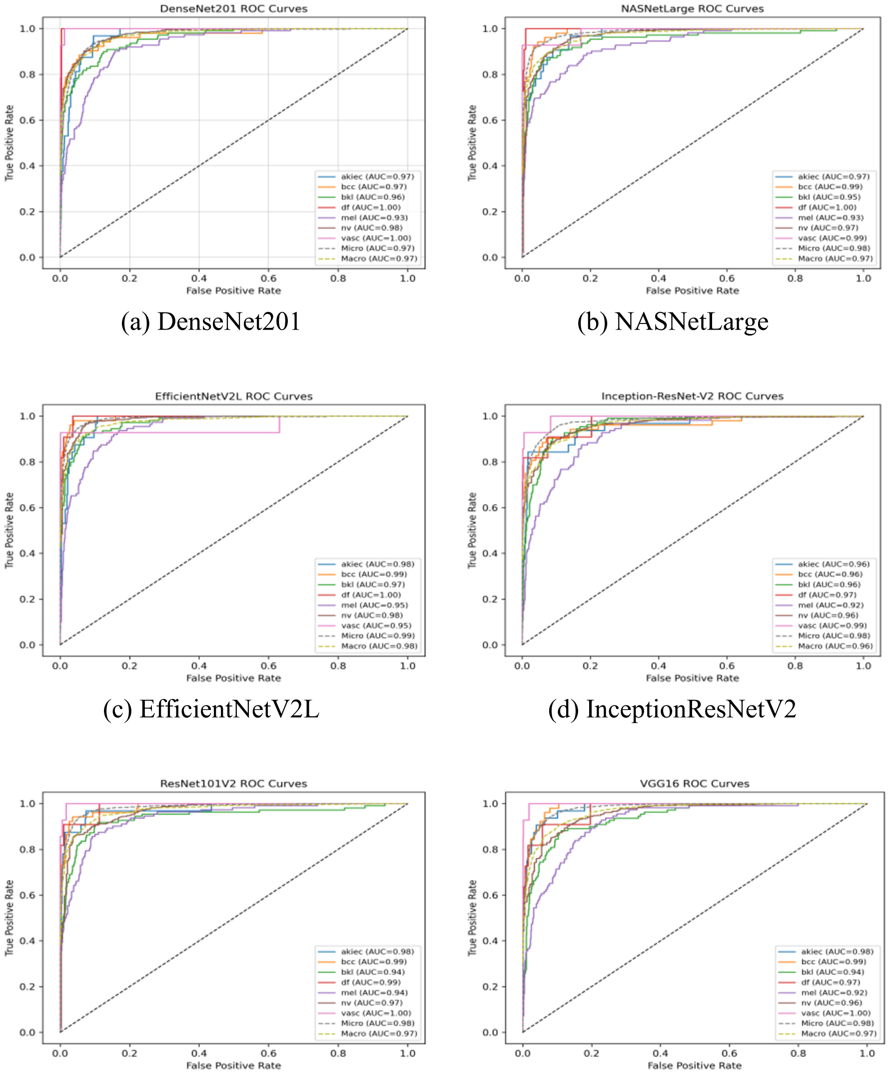


(e) ResNet101V2



(f) VGG16

Fig. 4. Confusion matrices of InceptionResNetV2, NASNetLarge, EfficientNetV2L, DenseNet201, ResNet101V2, and VGG16 evaluated on the seven-class skin disease test set.



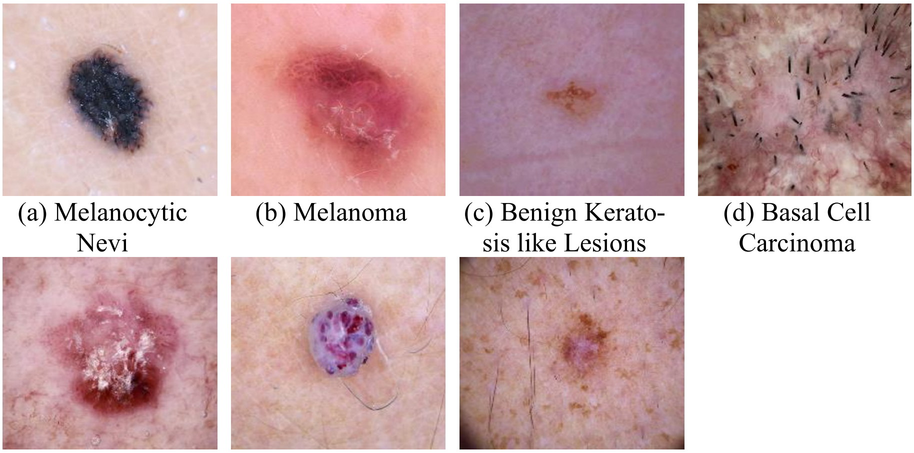
(a) DenseNet201 (b) NASNetLarge

(c) EfficientNetV2L (d) InceptionResNetV2

(e) ResNet101V2 (f) VGG16

Fig. 5. Receiver Operating Characteristic (ROC) curves with Micro-AUC and Macro-AUC scores for all transfer learning models on the test dataset.

Overall, EfficientNetV2L proved to be the most reliable model, offering the optimal balance of accuracy, generalization, and prediction consistency. The poor performance of older or less optimized architectures also highlights the importance of modern CNN architectures in multi-class skin disease classification. The above findings highlight the potential of EfficientNetV2L-based systems in automated dermatological diagnosis.



(a) Melanocytic Nevi (b) Melanoma (c) Benign Keratosis like Lesions (d) Basal Cell Carcinoma

(e) Actinic Keratoses (f) Vascular Lesions (g)Dermatofibroma

Fig. 6. Sample test images for seven skin disease classes.

Table 5. Class-wise prediction results of representative test images.

Figures	**Actual Class**	**Predicted Class**					
		DenseNet201	**NASNet Large**	**VGG16**	**ResNet101 V2**	**InceptionResNetV2**	**Effcient-NetV2L**
a	Melanocytic Nevi	Melanocytic Nevi	Vascular Lesions	Melanocytic Nevi	Melanocytic Nevi	Melanocytic Nevi	Melanocytic Nevi
b	Melanoma	Melanocytic Nevi	Melanocytic Nevi	Melanocytic Nevi	Dermato fibroma	Benign Keratosis like Lesions	Melanoma
c	Benign Keratosis like Lesions	Benign Keratosis-like Lesions	Melanocytic Nevi	Benign Keratosis like Lesions	Benign Keratosis like Lesions	Benign Keratosis like Lesions	Benign Keratosis like Lesions
d	Basal Cell Carcinoma	Benign Keratosis like Lesions	Benign Keratosis like Lesions	Benign Keratosis like Lesions	Actinic Keratoses	Benign Keratosis like Lesions	Basal Cell Carcinoma
e	Actinic Keratoses	Actinic Keratoses	Melanocytic Nevi	Actinic Keratoses	Actinic Keratoses	Actinic Keratoses	Actinic Keratoses
f	Vascular Lesions	Vascular Lesions	Vascular Lesions	Basal Cell Carcinoma	Vascular Lesions	Vascular Lesions	Vascular Lesions
g	Dermatofibroma	Melanoma	Dermato fibroma	Basal Cell Carcinoma	Dermato fibroma	Dermatofibroma	Dermato fibroma

5 Conclusion

In this study, the effectiveness of the proposed transfer learning-based approach based on the use of deep convolutional neural networks for the classification of skin diseases based on images from the HAM10000 dataset is investigated. In particular, the effectiveness of six popular and state-of-the-art deep convolutional neural networks is compared under the same conditions. The experiments were performed using various techniques such as data augmentation, class weighting, the use of the Focal Loss function, and fine-tuning to deal with the class imbalance problem. The experiments demonstrated the significant effectiveness of the fine-tuning technique in improving the model's performance. In particular, the experiments demonstrated the effectiveness of the EfficientNetV2L model in achieving the best accuracy, the lowest loss function value, and the best balance among precision, recall, and F1-score compared to the other tested models. In particular, the model has the best performance for all classes of skin disease, both common and rare. NASNetLarge model demonstrated a high level of performance compared to the other tested models. On the other hand, the VGG16 and DenseNet201 models demonstrated higher performance compared to the state-of-the-art models. The evaluation of the model's performance is based on the analysis of the confusion matrix and the use of the ROC-AUC curves to demonstrate the reliability of the proposed approach. The proposed approach has high potential to be used as a computer-aided diagnostic tool and can be used in resource-scarce environments. In the future, the model's reliability will be further improved using the Grad-CAM technique and by estimating the uncertainty to further improve the model's reliability.

References

1. Paul, S.S., Islam, M.J., Dhar, N.S., Kaysir, M.R.: DLC–TiO_2–Au-Coated D-Shaped SPR Biosensor for Cancer Cell Detection. IEEE Sens J. 25, 39802–39810 (2025). https://doi.org/10.1109/JSEN.2025.3609695
2. Pan, S.J., Yang, Q.: A Survey on Transfer Learning. IEEE Trans Knowl Data Eng. 22, 1345–1359 (2010). https://doi.org/10.1109/TKDE.2009.191
3. Thwin, S.M., Park, H.-S.: Skin Lesion Classification Using a Deep Ensemble Model. Applied Sciences. 14, 5599 (2024). https://doi.org/10.3390/app14135599
4. Alotaibi, A., AlSaeed, D.: Skin Cancer Detection Using Transfer Learning and Deep Attention Mechanisms. Diagnostics. 15, 99 (2025). https://doi.org/10.3390/diagnostics15010099
5. Rashid, J., Ishfaq, M., Ali, G., Saeed, M.R., Hussain, M., Alkhalifah, T., Alturise, F., Samand, N.: Skin Cancer Disease Detection Using Transfer Learning Technique. Applied Sciences. 12, 5714 (2022). https://doi.org/10.3390/app12115714
6. Zoph, B., Vasudevan, V., Shlens, J., Le, Q. V.: Learning Transferable Architectures for Scalable Image Recognition. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8697–8710. IEEE (2018)
7. Simonyan, K., Zisserman, A.: Very Deep Convolutional Networks for Large-Scale Image Recognition. (2015)
8. He, K., Zhang, X., Ren, S., Sun, J.: Identity Mappings in Deep Residual Networks. Presented at the (2016)

9. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely Connected Convolutional Networks. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2261–2269. IEEE (2017)
10. Tan, M., Le, Q. V.: EfficientNetV2: Smaller Models and Faster Training. (2021)