Explainable Deep Learning for Automated Skin Cancer Detection Using Advanced CNN Architectures on Dermoscopic Images

Adel Rajab
Computer Science Department-College of Computer Science and Information Systems, Najran University, Najran, 61441, Saudi Arabia

***Abstract*—Skin cancer is a considerable health issue worldwide, occurring when pigment cells turn malignant. However, diagnosing skin lesions is difficult for dermatologists because most lesions have similar characteristics. Initial detection is essential because it significantly increases the success rate of treatment and survival rates. In the past few decades, the rapid development of artificial intelligence has made it possible to build automated diagnostic systems based on large histopathology-validated image datasets. In this study, we introduce a deep learning solution for multi-class skin cancer classification based on state-of-the-art convolutional neural networks (CNNs) on the HAM10000+ISC image dataset. We used pre-trained CNN backbones, InceptionV3, DenseNet121, ResNet50, and VGG16, initialized with weights from ImageNet, for feature extraction, fine-tuning, and evaluation. Among the models, InceptionV3 achieved the highest accuracy of 76% and an ROC score of 0.967. To enhance interpretability, we used explainable AI (XAI) methods, Grad-CAM, Grad-CAM++, and class-wise attention maps, to examine both correctly and incorrectly classified images. The experiment demonstrates that the suggested system is not only characterized by high classification accuracy, but also by the ability to explain and visualize, which is a significant advantage for dermatologists when diagnosing skin cancer early and correctly.**

Keywords*—Deep learning models; skin cancer detection; image processing; Grad-CAM

I. Introduction

Recent breakthroughs in deep learning and computer vision have greatly impacted the medical diagnostic community, making it possible to develop accurate automated disease diagnosis systems [1]. Skin cancer is one of the numerous varieties of diseases observed to be one of the most common diseases, as well as a rapidly increasing cancer in the world. The American Cancer Society estimates that 2,041,910 cancer cases and 618,120 cancer deaths will be realized in 2025 in the United States. Although the death rate from cancer has been declining, the incidence and disparities in cancer highlight the need for better early detection, especially in high-risk cancers like melanoma [2]. It is important that skin lesions be detected early and properly classified to enhance better survival.

Conventional diagnosis techniques, like visual analysis and dermoscopy, are highly dependent on the expertise of dermatologists and are prone to inter-observer variability, thereby emphasizing the need for developing accurate computer-aided diagnosis systems. Deep learning has been recognized as a promising paradigm for medical image analysis due to its learning capability to extract hierarchical features from raw data [3].

Convolutional Neural Networks (CNNs) have been effective in the classification of skin cancer through their ability to detect both low-level and high-level features of the images [4]. VGG, ResNet, and DenseNet models have shown their efficiency in CNNs because of their ability to handle the depth of the network, residual learning, and improved gradient flow. However, despite the progress made, there are still some obstacles that need to be managed to achieve good performance. Therefore, this study proposes and examines the performance of VGG, ResNet, and DenseNet models for skin cancer classification [5], [6], [7].

Diagnosis of skin cancer is largely dependent on the visual inspection of the skin, which is prone to variability and subjectivity. Although promising outcomes have been obtained with deep learning models in terms of classifying skin lesions, high accuracy is still not reached with respect to general types of skin lesions. There is a need for a systematic comparison of the latest CNN models, such as VGG, ResNet, and DenseNet, to determine the best approach for accurate skin cancer detection. Furthermore, one of the biggest challenges that current deep learning models are facing is the lack of interpretability, which is still a major hindrance to the broad adoption of these models in clinical settings [8] [9].

Despite the progress made by deep learning techniques in improving the classification of skin lesions, there have been various limitations noted with the current automated techniques. For instance, various techniques have shown high performance using particular datasets, while their performance is questionable using other datasets. Furthermore, most CNN-based techniques have shown poor interpretability, thus making it hard to trust their predictions. Another limitation noted is class imbalance, where most data sets, such as HAM10000, have imbalanced classes, thus forcing the CNN models to bias towards certain types of lesions. To tackle the above-mentioned challenges, this study aims to explore various popular CNN models, such as VGG16, ResNet50, DenseNet121, and InceptionV3, for multi-class classification of skin lesions. Additionally, Explainable AI techniques, such as Grad-CAM,

Grad-CAM++, and class-specific attention, have been included to provide interpretability to CNN models, thus helping clinicians trust their predictions.

The primary objective of this research is to design and evaluate deep learning models for precise and automated skin cancer diagnosis using an advanced CNN. Additionally, to enhance the interpretability of the decision-making process, this study also integrates Explainable AI (XAI) methods, such as Grad-CAM, Grad-CAM++, and class-wise attention maps.

These explainability methods are applied to both correctly and incorrectly classified lesions to analyze how the models are making their predictions. These methods provide useful explanations in the form of visualizations that help to identify the specific areas of the lesion that are contributing to the model's predictions. This increased interpretability of the model's predictions helps to increase confidence in the reliability of the model for clinical use. Additionally, the use of XAI methods allows dermatologists to determine if the model is paying attention to the appropriate areas of the lesion and not just the background, which can be irrelevant.

The research contributes to the progress of a trustworthy deep learning system for the automated diagnosis of skin cancer. By comparing the VGG, ResNet, and DenseNet models, it determines efficient models that can help dermatologists in making an early diagnosis. The primary objectives of this study are as follows:

1) Leverage multiple methods in training and optimizing a multi-class skin lesion detection model with the latest convolutional neural networks, including VGG16, ResNet50, DenseNet121, and InceptionV3.

2) Conduct an extensive performance evaluation of the model using several evaluation measures like accuracy, precision, recall, and ROC AUC.

3) Use Explainable AI models which include Grad-CAM, Grad-CAM++, and class attention maps to interpret the model and build more trust in their model decision process.

Moreover, key contributions are listed below:

1) This study presents a comprehensive comparison of four popular convolutional neural networks in classifying skin lesions into multiple classes using the HAM10000 dataset.

2) This study proposes class weighting and data augmentation techniques to mitigate class imbalance in the dataset.

3) This study leverages Explainable AI techniques to increase model interpretability and trust in decision-making in dermatology and healthcare.

4) This study offers a paradigm that is both sensitive and explainable when it comes to model performance in making decisions in dermatology and healthcare.

The remainder of the study is organized in the following way: Section II locates related work or literature review. Section III contains discussions on the proposed study in terms of conducting some steps. The discussion of the key findings and interpretation is described in Section IV. Outcome or future directions are provided in Section V. Finally, Section VI concludes the study.

II. Literature Review

A. Hybrid Machine Learning and Deep Learning

Recent works suggest hybrid models of machine learning and deep learning for the detection of melanoma. A hybrid feature extraction method involving the best neural networks and handcrafted features reached 93% accuracy with recall values of 99.7% for benign and 86% for malignant samples, thus outperforming dermatologists and previous automated systems [10]. Similarly, in [11], the authors discuss the application of ML and DL models, particularly CNN, RNN, and transfer learning models for early detection of skin cancer. The SVM-PSO has a maximum accuracy of 97.5%.

B. Ensemble Deep Learning Models

The VGG, CapsNet, and ResNet architectures were trained to produce an ensemble model for skin cancer detection over the ISIC dataset. The accuracy of the individual models was 79%, 75%, and 69%, respectively, while the ensemble model reached 93.5% accuracy, thus outperforming individual deep learning models and traditional machine learning models [12]. The two ensemble models of VGG-16 + ResNet-50 and VGG-19 + Xception were trained on more than 3,000 skin images, showing 100% training and 85% testing accuracy. The ensemble models harnessed the best attributes of feature extraction, performing better than individual models, emphasizing improved reliability and accuracy for early skin cancer diagnosis [13]. The ensemble model of ResNet, EfficientNet, and VGG16, trained on the HAM10000 dataset, showed 99.1% accuracy for benign vs. malignant skin lesion classification. The rigorous preprocessing step improved image quality, and the ensemble model utilized the varied capabilities of feature representation, proving the efficiency of combining the best models for high accuracy [14].

C. Deep Learning Architectures for Skin Lesion Classification

Dermoscopic images and patients' metadata, such as age, gender, and body location, can be detected using the Inception-ResNet-v2 architecture. The model reached 89.3% accuracy for four-class classification and 94.5% accuracy for benign vs. malignant classification on 57,536 images, thus improving performance by at least 5% with the integration of metadata [15]. A CNN-based deep learning model with ESRGAN preprocessing was employed on the ISIC2018 dataset (3,533 images) for the classification of skin lesions. Transfer learning models (ResNet50, InceptionV3, Inception-ResNet) resulted in 83.7-85.8% accuracy, while the proposed CNN resulted in 83.2% accuracy, proving the effectiveness of image enhancement and preprocessing techniques for improved classification accuracy [16]. An attention-enhanced DenseNet-121 model with ICSOA hyperparameter tuning was employed on various datasets (ISIC 2017-2020, PH2, HAM10000) including UNet-based segmentation. The model resulted in classification accuracies of 99.55-99.87%, outperforming Insinet, N-DNN, and Skin-Net models, proving the effectiveness of attention mechanisms and optimal feature selection for skin lesion classification [17].

D. Fine-Tuned and Customized Deep learning Models

A fusion-level deep learning model integrating EfficientNet-B0, EfficientNet-B2, and ResNet50 was employed on the Kaggle Skin Diseases dataset (27,153 images) for ten-class skin disease classification. The model resulted in 99.14% accuracy with high precision, recall, and F1-score, proving its effectiveness for accurate classification with multi-architecture feature fusion [18]. The fine-tuned deep learning models, such as EfficientNetB0, ResNet34, VGG16, Inception_v3, and DenseNet-121, were further improved by adding more dense layers for skin cancer classification. The fine-tuned DenseNet-121 model showed 87% accuracy, which was better than the conventional machine learning models CNN, SVM, and Random Forest, emphasizing the need for customized models and layer optimization [19].

III. Methodology

This study has developed a deep learning based model that can help in the automatic classification of skin lesions based on dermoscopic images of the HAM10000 +ISIC dataset. The suggested framework is presented in Fig. 1 and has six stages. The first stage involves collecting the dataset, and to guarantee that images are consistent and high-quality, various image-processing techniques were applied. The second stage involves performing an exploratory data analysis to analyze the distribution of various types of skin lesions and handle class imbalance issues in the dataset. In the third stage, we divide our dataset using stratified sampling to ensure that we do not have any potential issues related to class imbalance in the dataset. In addition to this, we have also applied data augmentation techniques to ensure that our model is less prone to overfitting. For feature extraction and classification purposes, we have utilized different well-established pretrained CNN models. The efficiency of the anticipated model in terms of accuracy is ensured using comprehensive statistical evaluation metrics named ROC curve and Precision-Recall curve to understand the diagnostic potential of proposed model. In addition to this, this research have also utilized different visualization techniques named Grad-CAM and Grad-CAM++ to ensure transparency in our model's decision-making potential.

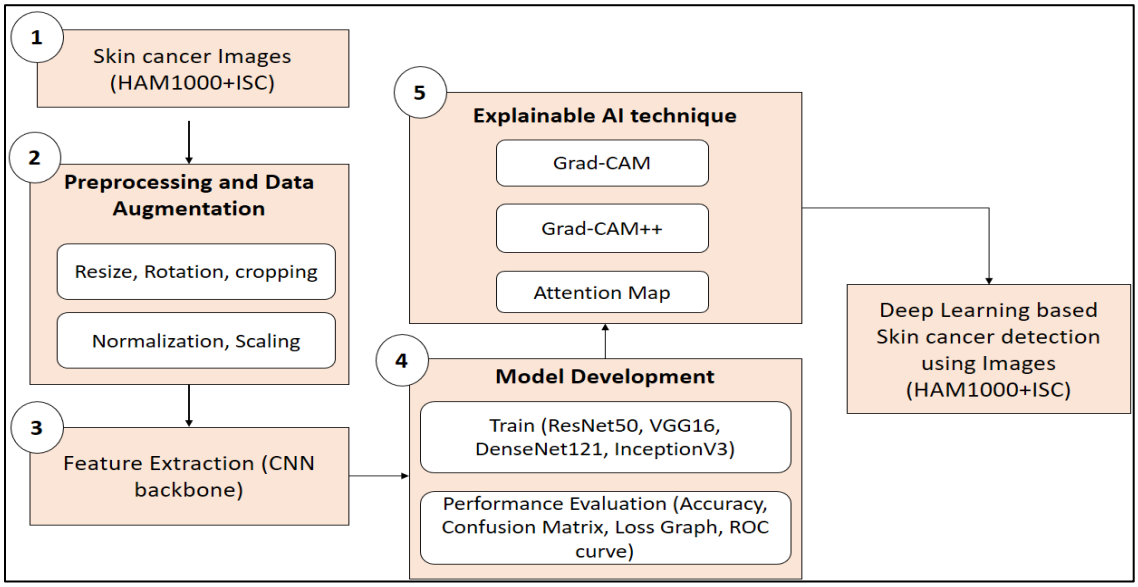

Fig. 1. Research methodology.

1) Dataset collection: The datastet used in this study was downloaded from Kaggle portal [20]. The datastet contains HANM1000 +ISIC images with seven classes, as represented in Fig. 2. The per-class samples and names are shown in Fig. 3.

The classes in the HAM10000 dataset have a significant class imbalance problem, meaning that some classes have a higher number of instances than others. In order to resolve this problem, a class weighting strategy was implemented in the training process of the model. The class weights were computed by utilizing a balanced class weight function from the scikit-learn library. In this way, more importance will be assigned to the minority classes in the training process, and the model will be less biased towards the majority classes in the dataset.

2) Data augmentation: For efficient computation, using ImageDataGenerator in TensorFlow/Keras, all images have been resized to 128 x 128 pixels and normalized between [0,1]. Furthermore to optimize GPU memory usage to speed up training process a batch size of 32 is used [21].

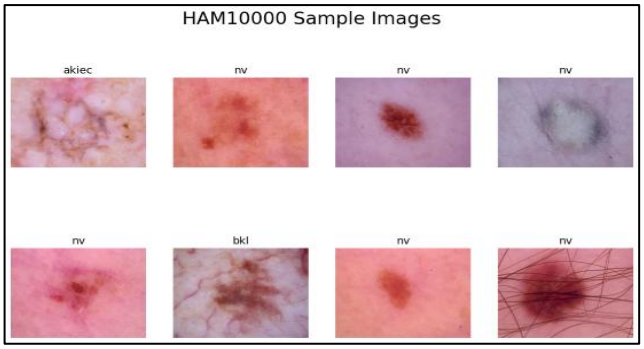

Fig. 2. Sample images of dataset.

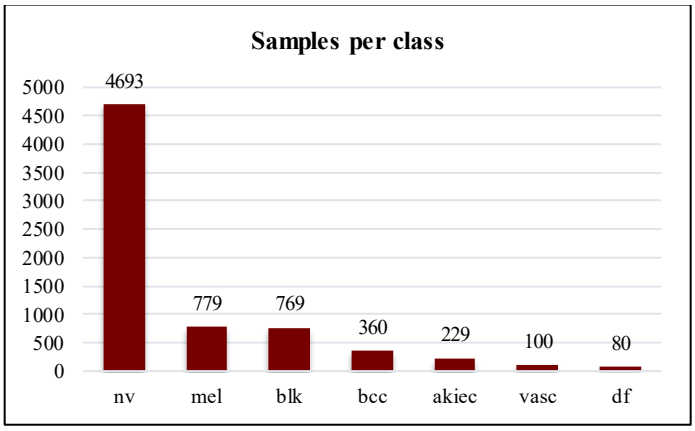

Fig. 3. Per class samples.

To assist the model in generalizing well and to account for the types of variations that are typically encountered in a clinical environment, extensive data augmentation strategies were used on the images used in training. Several types of transformations were randomly incorporated to create a diverse set of images. This included rotating images up to ±25 degrees, shifting images horizontally or vertically up to ±15%, zooming images up to ±25%, and applying a shear transformation of 0.1. Flipping images both horizontally and vertically was also used [22]. Class imbalance was addressed using class weighted loss strategy based on scikit's learn's balanced class weight function.

A. Feature Extraction

Feature extraction is a process through which images are transformed into a form that is easy to interpret, showing key features and trends in images. This process is achieved through convolutional neural networks in deep learning, which automatically learn features from images [23] [24] [25].

In this study, feature extraction was implemented through backbone CNN architecture:

$$\mathcal{F} = CNN_{backbone}\ (X) \quad (1)$$

where,

- $X \in \mathbb{R}^{\mathbb{H}\ x\ \mathbb{W}\ x3}$ is the input image
- $\mathcal{F} \in \mathbb{R}^{h\ x\ w\ x\ d}$ is the extracted feature map
- $\mathbb{H}, \mathbb{W}$ is the input image height and width (128 x 128)
- $h\ x\ w$ represents the spatial dimension of feature map
- d denotes the number of feature channel

The Global Average Pooling (GAP) layer converts spatial feature map $\mathcal{F}$ into a compact feature vector $f \in \mathbb{R}^d$:

$$f_j = \frac{1}{h.w}\sum_{i=1}^{h}\sum_{k=1}^{w}\mathcal{F}_{i,j,k} \quad \forall j = 1,2,3,\ldots,d \quad (2)$$

This vector is then forwarded to the fully connected layers for classification.

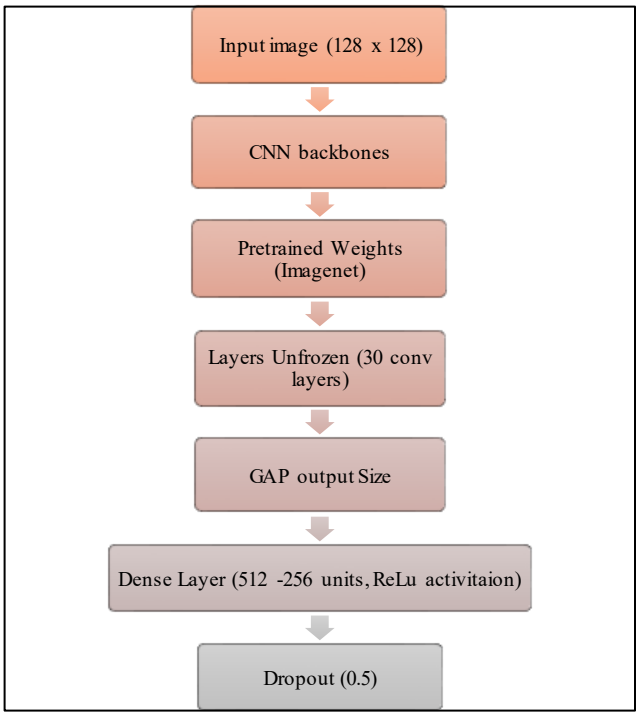

Fig. 4. Parameters used for feature extraction.

Fig. 4 represents parameters for CNN backbones to extract useful features for model training. The features learned by the network form a basis on which classification is made. The features allow the network to differentiate between classes in images. The features learned in this process are used to classify images into different classes. When dealing with images in a medical context, such as dermoscopic images, feature extraction is a key process in distinguishing between images based on their morphological differences. This enhances accuracy and generalization. The model uses transfer learning to classify images through a combination of ResNet50, DenseNet121, InceptionV3, and VGG16 networks, which have already been trained on a large number of images.

B. Model Development

In this study four advanced CNN architectures were employed as discussed below:

1) ResNet50: This model is a 50-layer residual network, where skip connections are incorporated to facilitate the flow of information across the layers. This helps to prevent the occurrence of the vanishing gradient problem when training deep networks [26]. Fig. 5 illustrates the various parameters of ResNet50 architecture.

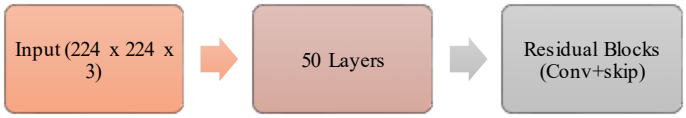

Fig. 5. ResNet50 parameters.

2) DenseNet121: This model features a series of connections where each layer is connected to the previous ones. This allows for the effective reuse of features across the network, which helps to improve the flow of information [27]. Fig. 6 represents the architectural specifications for DenseNet121 model.

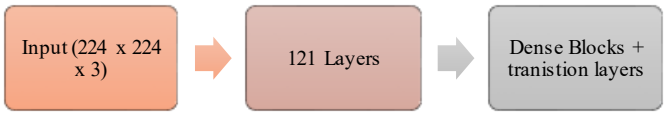

Fig. 6. Parameters for DenseNet121 architecture.

3) InceptionV3: This model features a multi-branch architecture where the image data is processed by a series of parallel convolutional filters. This helps to improve the features learned by the model by incorporating features of varying sizes [28]. Fig. 7 illustrates the parameters defined for InceptionV3 model for training.

Fig. 7. Architectural parameters of InceptionV3.

4) VGG16: This model features a 16-layer convolutional neural network known for its simplicity and consistency. This model features small 3×3 filters used for convolutional processing. Despite its simplicity, this model has proven to be effective for image classification [29]. Fig. 8 shows the parameters such as input image size, total number of layers and maxpooling block.

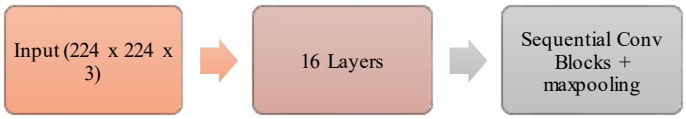

Fig. 8. Parameters of VGG16 model.

C. Explainability and Artificial Intelligence

Explainability in deep learning techniques is a set of techniques used to aid in the understanding of the decision-making process of the model. In image classification, it is often done with visual attention maps, which include techniques such as Grad-CAM and Grad-CAM++. It helps to build confidence in the clinical diagnosis by providing insight into the basis of the model's prediction of the type of lesion. Furthermore confirms the model is attending to the relevant parts of the image, rather than the irrelevant background [30]. In this study, the following techniques were employed:

1) Grad-CAM: Gradient-weighted Class Activation Mapping (Grad-CAM) was utilized to identify areas in each image that had a major influence on the predictions made by the model. The model focused on the final convolutional layer to create a heatmap highlighting important features in each image.

2) Grad-CAM++: To obtain a precise localization of features, particularly in images containing small lesions or multiple areas of interest, Grad-CAM++ was utilized. This model provides a sharper and more detailed attention map compared to Grad-CAM.

3) Class-wise attention maps: The average attention maps were generated for each class of lesions by summing up the output from multiple images using Grad-CAM++. This provided a deeper understanding of features used by the model to differentiate between different types of lesions [31].

IV. Results and Discussion

The experiment has been accompanied using Kaggle Notebook, and the Keras library of Python's deep learning. The model has been trained using a batch size of 32, and 40 epochs. Adam optimization has been used in the experiment, and the learning rate has been set to 1×10^{-4} to ensure the efficiency of the model. A categorical cross-entropy has been used as the loss function, as it is the most appropriate function for the classification problem. In an effort to enhance stability in the training process and prevent overfitting in the model, several regularization techniques were added to the model. These included the use of Dropout layers, Batch Normalization, as well as an adaptive learning rate adjustment and early stopping. The learning rate adjustment was achieved with the ReduceLROnPlateau callback function, which reduced the learning rate by a factor of 0.3 if there was no enhancement in the validation loss for three consecutive epochs. Additionally, an EarlyStopping method was used with a patience of seven epochs to stop the training process if there were no improvements in the validation loss.

Table I shows the division of training testing and validation samples. Specifically, that dataset was spilt into 70% training, 20% testing, and 10% validation using the train_test_split function with a fixed random seed (random state =42)

TABLE. I. Samples Division

Training	Testing	Validation
7010	2003	1002

A. Performance Comparison of Deep Learning Models

Table II presents a summary of the performance of several deep learning models on standard evaluation parameters. Among these models, InceptionV3 performed the best, with an accuracy of 76%, recall of 77%, and F1 score of 70%, although its precision was marginally lower at 66%. The second-best model was DenseNet121, which reported an accuracy of 74%, precision of 72%, recall of 71%, and F1 score of 71%. ResNet50 performed moderately well, with an accuracy of 71%, precision of 68%, recall of 65%, and F1 score of 65%. VGG16, on the other hand, performed the worst, with an accuracy of 67%, precision of 60%, recall of 65%, and F1 score of 60%. The superior performance of InceptionV3 can be attributed to its architecture, which incorporates factorized convolutions and inception modules to effectively extract multi-scale features, thereby leading to higher accuracy and

recall. The dense connectivity of DenseNet121 helps it perform better, as it enables effective feature propagation and prevents vanishing gradients. Although ResNet50 is a capable model, it lags slightly behind InceptionV3 and DenseNet121, possibly because of the relatively less extensive use of feature reuse. The poor performance of VGG16 emphasizes the weakness of its simple and uniform architecture in handling complex hierarchical features.

TABLE. II. PERFORMANCE EVALUATION OF DL MODELS

DL models	Accuracy	Precision	Recall	F1 score
ResNet50	71%	68%	65%	65%
DenseNet121	74%	72%	71%	71%
InceptionV3	76%	66%	77%	70%
VGG16	67%	60%	65%	60%

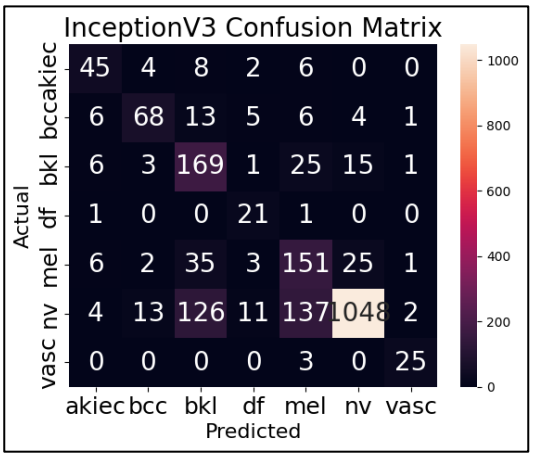

Fig. 9. Confusion matrix for the InceptionV3 model.

Fig. 9 illustrates the performance of the InceptionV3 model over the unseen test dataset. Out of the total samples (2003), the model correctly classified 1527 with 476 misclassifications.

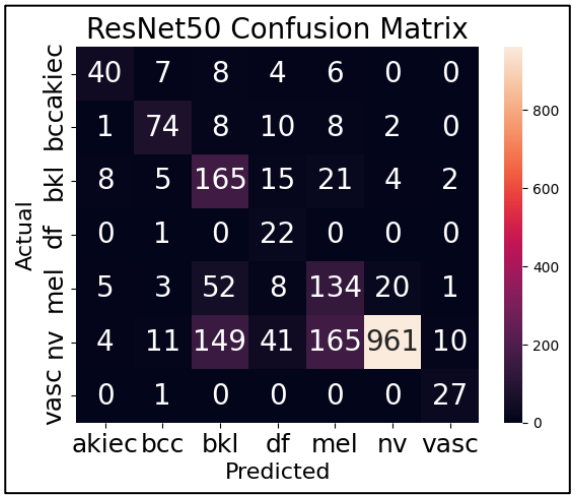

Fig. 10. Confusion matrix for the ResNet50.

Fig. 10 illustrates the performance of the ResNet50 model, where the model accurately predicted 1423 samples while 580 samples were misclassified.

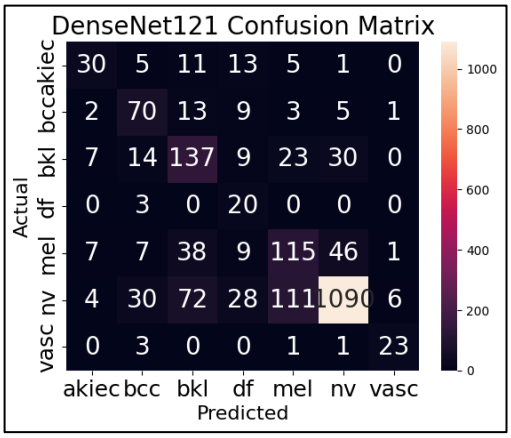

Fig. 11. Confusion matrix for DenseNet121.

Fig. 11 shows the performance of the DenseNet121 over test samples. The model rightly predicted 1463 samples from the total sample (2003), while 540 samples were misclassified.

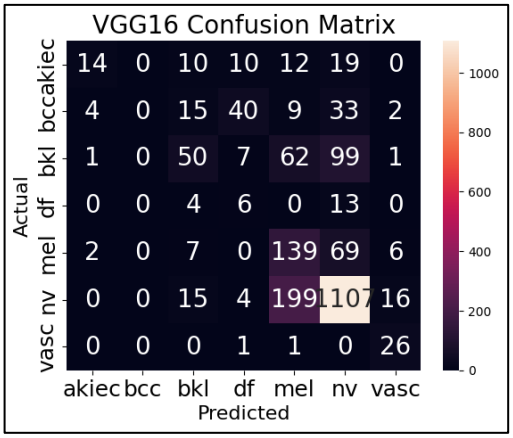

Fig. 12. Confusion matrix for the VGG16 model.

Fig. 12 shows the confusion matrix for the VGG16 model, where out of total samples (2003) the model correctly classified 1342 samples with 661 misclassifications.

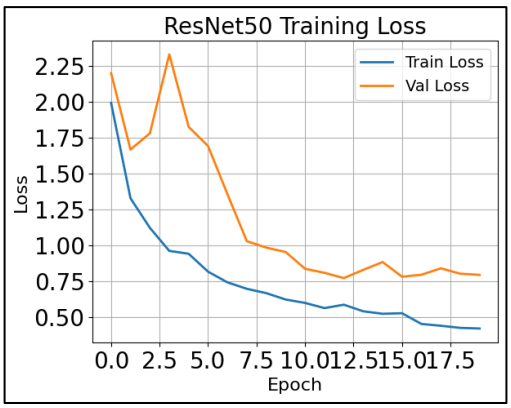

Fig. 13. Loss curve for ResNet50 model.

Fig. 13 represents the train and validation loss. The orange curve depicts the validation loss while blue curve shows the train loss. Both losses decrease significantly during initial epochs but minor fluctuations were observed in validation loss,

around at epoch 5 the validation loss suddenly rises at peak and then drops. The initial drop in both training and validation loss suggests that the model is effectively learning relevant and discriminative features from the data. The slight increase in validation loss during the training process may be due to overfitting or slight oscillations in the loss due to batch-wise changes during the optimization process. However, as the validation loss eventually drops and stabilizes, it is a clear indication that the model has successfully regained its capacity to generalize. This is probably achieved through various regularization techniques such as dropout and learning rate schedules.

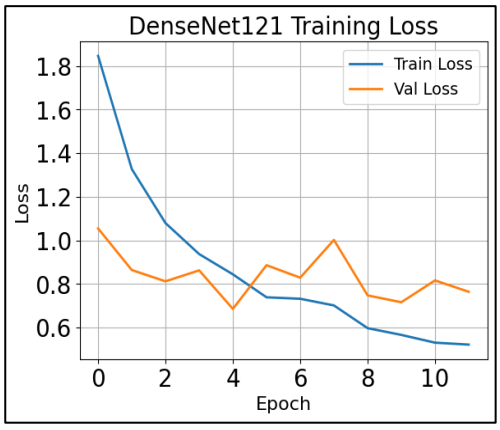

Fig. 14. Loss graphs for DenseNet 121.

Fig. 14 shows the train and validation loss for the DenseNet121 model. The train loss decreases from a loss value of 1.8 during initial epochs, while the validation continuously fluctuates. When reaching at epoch 4, the validation loss overlaps train loss. The gradual decrease in the training loss indicates that the optimization is going on well and the gradients are flowing well in the network, which can be credited to the dense connections in the DenseNet121 model. At around epoch 4, the training and validation loss curves are almost the same, which indicates that the model is generalizing well at this stage.

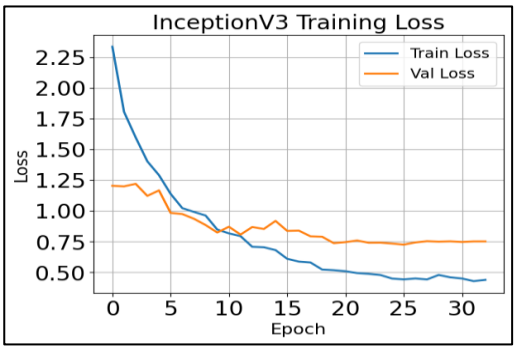

Fig. 15. Loss curves for InceptionV3 model.

Fig. 15 shows the train and valid loss generated by InceptionV3 model. During initial epochs, the train loss decreases, however, when reaching around epoch, 10 minor fluctuations were observed. The validation drops from a loss value of 1.25, but continuously fluctuates and remained within loss value (1). The steady reduction in the training loss over the first few epochs signifies that the InceptionV3 model is effectively learning features and converging. The minor oscillations in both the training and validation loss can be attributed to batch-wise differences and the model's adjustment to the complexity of multi-class lesion features. The fact that the validation loss is more or less steady at 1 indicates that the model is maintaining its generalization capabilities well and not overfitting much, which is probably aided by regularization techniques and the inception modules in the model that are efficiently extracting multi-scale information from the lesions.

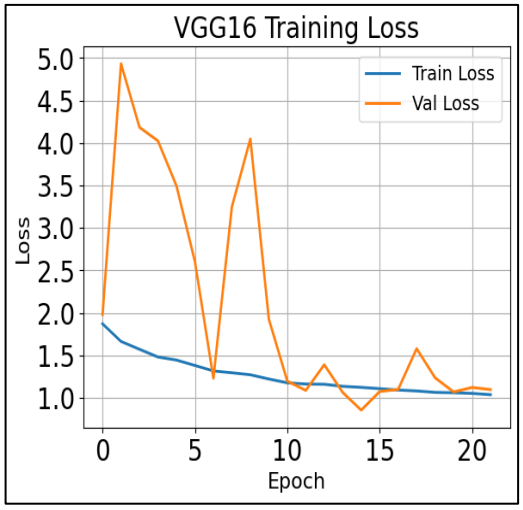

Fig. 16. Loss curve for the VGG16 model.

The loss curve of training and validation of the VGG16 model is shown in Fig. 16. Its loss during the training decreases gradually, starting with a level of about 2, showing that the model is learning the feature representations. However, the validation loss varies erratically, peaking at 5 and showing periodic peaks and troughs at around epoch 5 and later epochs. The erratic validation loss shows that VGG16 is not generalizing well on this data. This is because VGG16 is a shallow and sequential model, which is not good at representing complex multi-scale features compared to other models like ResNet50 or InceptionV3. The periodic peaks in the validation loss could also be due to batch variations and early stages of overfitting. However, the steady drop in the training loss shows that the model is still learning good features, even if it is not generalizing well.

B. Explainability Analysis Using Grad-CAM++

Fig. 17 shows Grad-CAM++ results for two dermoscopic images of melanocytic nevi (nv). For the left image, the prediction confidence of the model was relatively low, at 0.33, and the Grad-CAM++ results show broad, diffuse areas in the image and the surrounding skin. On the other hand, the Grad-CAM++ results for the right image show that the prediction confidence was relatively high, at 0.77, and the attention was focused on the lesion, demonstrating that the model was successful in identifying the most discriminative features. The Grad-CAM++ results are based on the computation of the weighted sum of the positive gradients in the last convolutional layer, resulting in class-specific localization maps that show the areas that the CNN uses to make the prediction. The difference in the attention distribution in the Grad-CAM++ results for the

two images demonstrates the potential of Grad-CAM++ to show the confidence of the CNN in its prediction and to provide insights into the CNN's prediction process for the lesion regions in the images.

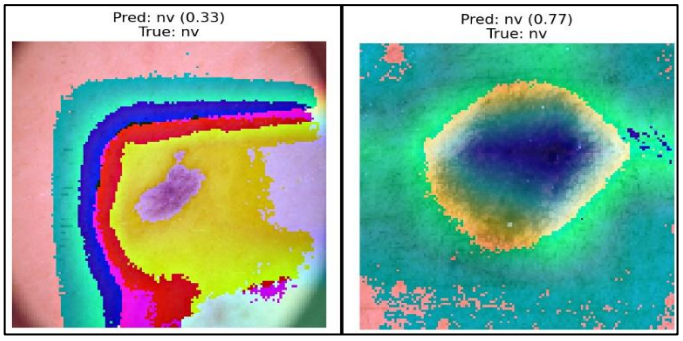

Fig. 17. Correct classification.

Fig. 18 shows Grad-CAM++ attention maps for two dermoscopic images that the model incorrectly classified. On the left, the model incorrectly classified the dermoscopic image as melanoma (mel, 0.49) when in fact it should have been classified as melanocytic nevus (nv). On the right, the model incorrectly classified the dermoscopic image as a benign keratosis-like lesion (bkl, 0.73) when in fact it should have been classified as melanoma (mel).

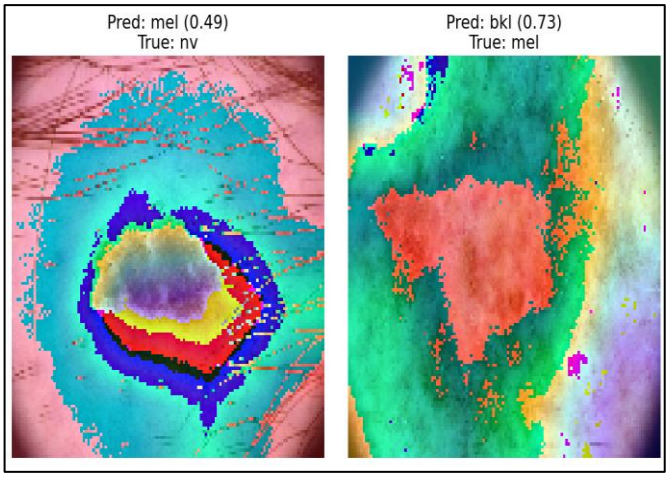

Fig. 18. Missed classification.

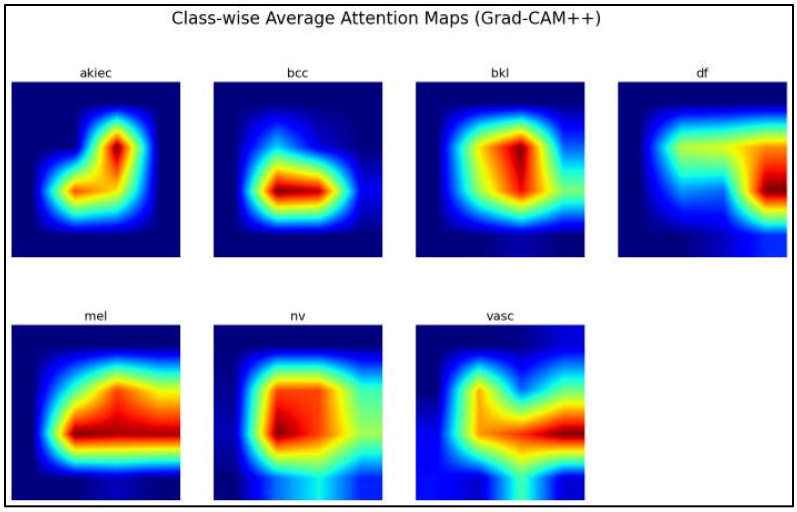

Fig. 19. Class-wise average attention maps.

Fig. 19 shows the class-wise average attention maps obtained using Grad-CAM++ for seven classes of skin lesions: akiec, bcc, bkl, df, mel, nv, and vasc. These attention maps help identify the regions of interest in the input images that the CNN uses for prediction in each class. The hotter colors, such as red and yellow, in the attention maps represent regions of high attention, while the cooler colors, such as blue, represent regions of low attention in the input image. For instance, the attention maps for mel and nv classes have elongated attention regions in the central part of the lesions, indicating that the model is considering larger regions of the input image for making predictions. On the other hand, the attention maps for bcc and akiec classes are more focused, indicating that predictions for these classes are made using localized features. Grad-CAM++ is an improvement over Grad-CAM because it uses higher-order gradients. This improvement enables the CNN to better focus on localized regions of importance, even if there are multiple instances or fine details in the image. Consequently, the CNN can better identify the importance of the features that are driving its prediction.

C. Comparative Analysis

Comparison of the performance of models is one of the objectives of this study to select the best-fit model.

Fig. 20 shows the ROC curve analysis for all the models that were evaluated. The ROC curve analyze the true positive rate against the false positive rate for different thresholds. The AUC value is a measure of the model's capability to assign higher scores to positive instances than to negative instances InceptionV3 had the highest AUC value of 0.967, which is the best discriminative capability. DenseNet121 had the second-best discriminative capability with an AUC value of 0.962, followed by ResNet50 with an AUC value of 0.956 and VGG16 with an AUC value of 0.938. InceptionV3's high AUC value can be attributed to its deeper architecture and inception modules that are efficient in capturing multi-scale features and

improving class separation. DenseNet121's good performance can be attributed to its dense connections that help in improving the flow of gradients and reusing features. Although ResNet50 and VGG16 are still good models, their lower AUC values imply slightly lower discriminative capabilities.

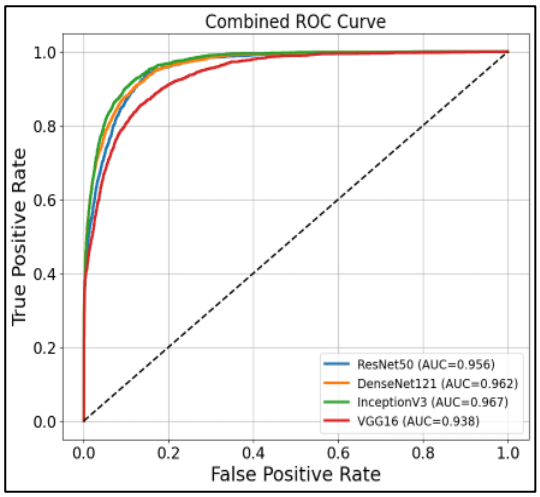

Fig. 20. ROC curve for all models.

Fig. 21 shows the comparison of all models based on their accuracy. InceptionV3 had the highest accuracy of 76%, followed by DenseNet121 with an accuracy of 74%, and then ResNet50 with an accuracy of 71%. The lowest accuracy was shown by VGG16, which only managed 67% accuracy. The differences in accuracy rates indicate the pros and cons of each model's architecture and its ability to extract features. InceptionV3 has the advantage of inception modules that apply multi-scale convolutional filters to effectively capture a broad range of lesion patterns.

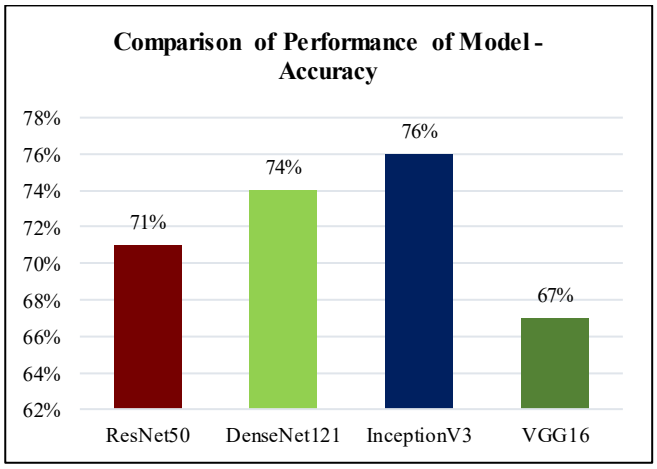

Fig. 21. Comparative analysis.

V. Discussion

Table II presents a summary of the performance of the tested CNN architectures on the various evaluation metrics. Among the architectures, InceptionV3 recorded the highest accuracy of 76%, recall of 77%, and F1-score of 70%, albeit a slightly lower precision of 66%. The second best architecture was DenseNet121, which recorded an accuracy of 74%, precision of 72%, recall of 71%, and F1-score of 71%. ResNet50 recorded a moderate performance with an accuracy of 71% and F1-score of 65%, while VGG16 recorded the lowest performance on all metrics, with an accuracy of 67% and an F1-score of 60%. The high performance of InceptionV3 is due to its factorized convolutions and inception modules, which are highly efficient in capturing multi-scale features and improving discriminative capabilities. The high performance of DenseNet121 is due to its dense connectivity, which enables feature reuse and prevents vanishing gradients. Although ResNet50 performed well, it is slightly less effective due to limited feature propagation, while VGG16's shallow and uniform architecture is highly ineffective in modeling complex hierarchical features, hence its low performance.

The best performance was obtained by InceptionV3, which can be attributed to its inception modules and factorized convolutions that help it to extract features at varying scales in skin lesions. DenseNet121 also performed well, owing to its dense connections that facilitate feature reuse and help gradients flow more freely through the network. ResNet50 produced moderate results, as it has limited feature propagation compared to DenseNet121. VGG16 produced the poorest results, as it has a limited depth compared to other models, making it difficult to extract features at varying scales in skin lesions.

In the past few decades, advanced deep learning models were explored for skin cancer detection by utilizing models such as EfficientNet, Vision Transformers, as well as hybrid models using CNNs and Transformers. Most of these studies have achieved classification accuracy rates of more than 90% using the HAM10000 dataset. In contrast, the present study compares various architectures. Another important aspect of the current study is the focus on the interpretation of the models' predictions, using the Grad-CAM and Grad-CAM++ methods, which is an important consideration in terms of trust in the models' predictions, especially in a clinical setting.

The confusion matrices (Fig. 9 to Fig. 12) depicts the performance of models over testing samples. InceptionV3 accurately classified 1,527 out of 2,003 test samples with 476 misclassifications (Fig. 9), indicating high discriminative ability. ResNet50 and DenseNet121 accurately classified 1,463 (540 misclassified) and 1,423 (580 misclassified) samples, respectively (Fig. 10- Fig. 11), while VGG16 accurately classified only 1,342 samples with 661 misclassifications (Fig. 12), indicating its poor generalization and feature representation capabilities.

The training and validation loss plots (Fig. 13 to Fig. 16) further confirm performance. For ResNet50 (Fig. 13), both training and validation losses started decreasing with minor fluctuations, indicating efficient convergence and generalization, facilitated by dropout and learning rate scheduling. For DenseNet121 (Fig. 14), the training loss remained stable, while the validation loss overlapped around epoch 4, indicating high generalization ability due to dense connections. For InceptionV3 (Fig. 15), the training loss remained stable with minor fluctuations in validation loss, indicating efficient multi-scale feature extraction and generalization. In contrast, VGG16 (Fig. 16) showed unstable

validation loss with frequent peaks, indicating high overfitting and poor generalization, as expected due to its shallow architecture.

The training and validation loss curve provides valuable insight into the performance of the models according to their ability to generalize. The low difference in the training and validation losses indicates that there was little room for overfitting and that the training process was stable. Moreover, to improve the robustness of models, data augmentation, early stoppage, and learning rate have been employed.

The integration of (XAI) techniques further helps to understand the interpretability of the models. The Grad-CAM++ attention maps (Fig. 17 to Fig. 19) demonstrate that the models are able to precisely attend to lesion-specific areas for correct predictions, while the misclassified images demonstrate general or incorrect attention patterns, pointing to areas of failure of the model. Class-wise average attention maps (Fig. 19) indicate the attention patterns vary across lesion types, with broader regions for melanocytic nevi (nv) and melanoma (mel), localized regions for basal cell carcinoma (bcc) and actinic keratoses (akiec). These results clearly show the ability of XAI techniques to improve clinical trust by providing visual explanations of the model's predictions.

Grad-CAM and Grad-CAM++ visualization also supported this, as it indicated that the models that produced better results paid more attention to the most important parts of the lesions, which indicates a strong correlation between model architecture, feature extraction, and accuracy.

ROC curve analysis (Fig. 20) verifies the discriminative power of the models. InceptionV3 had the highest AUC value of 0.967, followed by DenseNet121 (0.962), ResNet50 (0.956), and VGG16 (0.938), which emphasizes the superiority of deeper models and multi-scale feature extraction for discrimination. Accuracy comparison analysis (Fig. 21) also validates the same, ranking InceptionV3 as the best-performing model, followed by DenseNet121 and ResNet50, while VGG16 is the worst-performing model.

These findings collectively stress the importance of model architecture in both predictive accuracy and generalization. The integration of XAI methods enables the extraction of clinically important features, improving interpretability and facilitating the successful implementation of skin cancer diagnostic systems using deep learning models.

A. Practical Implications

The outcomes of this research have several key implications for practical applications in the field of dermatology. Firstly, the potential exists for high-performing models such as InceptionV3 and DenseNet121 to serve as a trustworthy decision-support system for dermatologists, enabling early detection and correct classification of skin lesions while possibly minimizing the occurrence of diagnostic mistakes. Secondly, the application of Explainable AI methods enables visual interpretation of model predictions, enabling dermatologists to confirm whether the model is concentrating on medically significant skin lesion areas instead of unnecessary artifacts, which could enhance AI-assisted decision-making.

B. Limitations

Nevertheless, the study also has some limitations. The models were trained and tested mainly on the HAM10000 dataset, which could limit their ability to generalize to images from other populations, devices, or settings. Moreover, the approach is based solely on dermoscopic images and excludes patient-specific clinical data, such as age, gender, or lesion site, which could further improve diagnostic performance and enable more accurate diagnostic model.

However, it should be noted that although this work has analyzed the application of transfer learning with various architectures for CNNs, more recent architectures such as transformer models or hybrid models combining convolutional and transformer architectures were not included in this study. Another limitation is that although this study has proposed models that were implemented and tested in a controlled experimental environment, it should be noted that these models have not yet been tested in a real-world clinical environment. Other potential improvements may come from future experiments that include more recent and more powerful deep learning architectures to improve diagnostic accuracy.

VI. Conclusion

This study presented a novel deep learning approach for multi-class skin lesion classification on the HAM10000 dataset. Various advanced CNN models, such as InceptionV3, DenseNet121, ResNet50, and VGG16, were fine-tuned and compared using a comprehensive comparative study. The InceptionV3 performed achieved the highest accuracy of 76% and an ROC-AUC value of 0.967. While the DenseNet121 and ResNet50 attained moderate performance, however, VGG16 performed relatively poor. The results clearly indicate that deeper networks with multi-scale feature extraction and enhanced gradient flow are more adept at handling complex lesion characteristics.

In addition to evaluating performance, this study also tackled the critical aspect of interpretability by incorporating (XAI) approaches, such as Grad-CAM, Grad-CAM++, and class-wise attention maps. These approaches allowed for the visual interpretation of the decision-making process of the models by pointing out the regions of the lesions that contributed to the classification results. This approach also enables to identify the areas of the skin lesions that were contributing the most to the predictions made by the models, which was important in the early diagnosis and accurate classification of skin cancer cases.

Collectively, these results illustrate that the integration of advanced CNN models with interpretability methods can lead to the development of a strong and practical automated system for skin cancer diagnosis.

Future studies would evaluate model on larger datasets that include patients from different ethnic backgrounds. This would not only help in validating the system's generalization capabilities but also increase confidence in its effectiveness in a real-world environment. By using both visual and clinical information through multimodal deep learning approaches, it may be possible to develop a more comprehensive and informed decision-making system.

Data Availability Statement

The dataset used in this study for early detection is publicly available at https://www.kaggle.com/datasets/nour12347653/skin-disease-detection-dataset-ham10000-isic

References

[1] M. Dildar et al., "Skin Cancer Detection: A Review Using Deep Learning Techniques," International Journal of Environmental Research and Public Health, vol. 18, no. 10, p. 5479, Jan. 2021, doi: 10.3390/ijerph18105479.

[2] R. L. Siegel, T. B. Kratzer, A. N. Giaquinto, H. Sung, and A. Jemal, "Cancer statistics, 2025," CA: A Cancer Journal for Clinicians, vol. 75, no. 1, pp. 10–45, 2025, doi: 10.3322/caac.21871.

[3] B. Sreedhar, M. Swamy B.E, and M. S. Kumar, "A Comparative Study of Melanoma Skin Cancer Detection in Traditional and Current Image Processing Techniques," in 2020 Fourth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC), Oct. 2020, pp. 654–658. doi: 10.1109/I-SMAC49090.2020.9243501.

[4] L. I. Mampitiya, N. Rathnayake, and S. D. Silva, "Efficient and Low-Cost Skin Cancer Detection System Implementation with a Comparative Study Between Traditional and CNN-Based Models," Journal of Computational and Cognitive Engineering, vol. 2, no. 3, pp. 226–235, 2023, doi: 10.47852/bonviewJCCE2202482.

[5] N. Girdhar, A. Sinha, and S. Gupta, "RETRACTED ARTICLE: DenseNet-II: an improved deep convolutional neural network for melanoma cancer detection," Soft Comput, vol. 27, no. 18, pp. 13285–13304, Sep. 2023, doi: 10.1007/s00500-022-07406-z.

[6] S. D and S. J, "Skin Cancer Detection using Convolutional Neural Network Models: VGG-16, VGG-19, ResNet and DenseNet," in 2024 5th International Conference on Data Intelligence and Cognitive Informatics (ICDICI), Nov. 2024, pp. 1254–1258. doi: 10.1109/ICDICI62993.2024.10810998.

[7] R. S. Sonawane, "Skin-Cancer Classification Using Deep Learning with DenseNet and VGG with Streamlit-Framework Implementation," masters, Dublin, National College of Ireland, 2023. Accessed: Feb. 22, 2026. [Online]. Available: https://norma.ncirl.ie/6669/

[8] P. Ahmed Abdalla et al., "TRANSFER LEARNING MODELS COMPARISON FOR DETECTING AND DIAGNOSING SKIN CANCER," Acta inform. Malays., vol. 7, no. 1, pp. 01–07, 2023, doi: 10.26480/aim.01.2023.01.07.

[9] A. A. Adegun and S. Viriri, "FCN-Based DenseNet Framework for Automated Detection and Classification of Skin Lesions in Dermoscopy Images," IEEE Access, vol. 8, pp. 150377–150396, 2020, doi: 10.1109/ACCESS.2020.3016651.

[10] J. V. Tembhurne, N. Hebbar, H. Y. Patil, and T. Diwan, "Skin cancer detection using ensemble of machine learning and deep learning techniques," Multimed Tools Appl, vol. 82, no. 18, pp. 27501–27524, Jul. 2023, doi: 10.1007/s11042-023-14697-3.

[11] S. Adamu et al., "The future of skin cancer diagnosis: a comprehensive systematic literature review of machine learning and deep learning models," Cogent Engineering, vol. 11, no. 1, p. 2395425, Dec. 2024, doi: 10.1080/23311916.2024.2395425.

[12] A. Imran, A. Nasir, M. Bilal, G. Sun, A. Alzahrani, and A. Almuhaimeed, "Skin Cancer Detection Using Combined Decision of Deep Learners," IEEE Access, vol. 10, pp. 118198–118212, 2022, doi: 10.1109/ACCESS.2022.3220329.

[13] A. Alshehri, "Skin-NeT: Skin Cancer Diagnosis Using VGG and ResNet-Based Ensemble Learning Approaches.," Traitement du Signal, vol. 41, no. 4, p. 1689, Aug. 2024, doi: 10.18280/ts.410405.

[14] S. Thakur and S. Sharma, "Ensemble Fusion: Skin Cancer Detection using ResNet, EfficientNet, and VGG Architectures," in 2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT), Jun. 2024, pp. 1–8. doi: 10.1109/ICCCNT61001.2024.10726235.

[15] R. Ahmadi Mehr and A. Ameri, "Skin Cancer Detection Based on Deep Learning," J Biomed Phys Eng, vol. 12, no. 6, pp. 559–568, Dec. 2022, doi: 10.31661/jbpe.v0i0.2207-1517.

[16] W. Gouda, N. U. Sama, G. Al-Waakid, M. Humayun, and N. Z. Jhanjhi, "Detection of Skin Cancer Based on Skin Lesion Images Using Deep Learning," Healthcare, vol. 10, no. 7, p. 1183, Jul. 2022, doi: 10.3390/healthcare10071183.

[17] N. N. Gajbhiye, S. R. Kadu, and S. Sadruddin, "An Efficient Skin Cancer Detection and Classification Using DenseNet-121 with Attention Mechanism.," International Journal of Intelligent Engineering & Systems, vol. 18, no. 5, p. 745, May 2025, doi: 10.22266/ijies2025.0630.52.

[18] M. Alruwaili and M. Mohamed, "An Integrated Deep Learning Model with EfficientNet and ResNet for Accurate Multi-Class Skin Disease Classification," Diagnostics, vol. 15, no. 5, p. 551, Jan. 2025, doi: 10.3390/diagnostics15050551.

[19] A. Bello, S.-C. Ng, and M.-F. Leung, "Skin Cancer Classification Using Fine-Tuned Transfer Learning of DENSENET-121," Applied Sciences, vol. 14, no. 17, p. 7707, Jan. 2024, doi: 10.3390/app14177707.

[20] "Skin Disease Detection Dataset (HAM10000 + ISIC)." Accessed: Feb. 21, 2026. [Online]. Available: https://www.kaggle.com/datasets/nour12347653/skin-disease-detection-dataset-ham10000-isic

[21] Z. Hussain, F. Gimenez, D. Yi, and D. Rubin, "Differential Data Augmentation Techniques for Medical Imaging Classification Tasks," AMIA Annu Symp Proc, vol. 2017, pp. 979–984, Apr. 2018.

[22] E. Goceri, "Medical image data augmentation: techniques, comparisons and interpretations," Artif Intell Rev, vol. 56, no. 11, pp. 12561–12605, Nov. 2023, doi: 10.1007/s10462-023-10453-z.

[23] S. K. Sundararajan, B. Sankaragomathi, and D. S. Priya, "Deep Belief CNN Feature Representation Based Content Based Image Retrieval for Medical Images," J Med Syst, vol. 43, no. 6, p. 174, May 2019, doi: 10.1007/s10916-019-1305-6.

[24] A. Golkarieh, K. Kiashemshaki, S. R. Boroujeni, and N. A. Isakan, "Advanced U-Net Architectures with CNN Backbones for Automated Lung Cancer Detection and Segmentation in Chest CT Images," Jul. 22, 2025, arXiv: arXiv:2507.09898. doi: 10.48550/arXiv.2507.09898.

[25] C. C. Atabansi et al., "E-TransConvNet: An enhanced transformer and convolutional network for medical image segmentation from ultrasound and CT images," Expert Systems with Applications, vol. 285, p. 128022, Aug. 2025, doi: 10.1016/j.eswa.2025.128022.

[26] "Detection of cancer from histopathology medical image data using ML with CNN ResNet-50 architecture," in Computational Intelligence in Healthcare Applications, Academic Press, 2022, pp. 237–254. doi: 10.1016/B978-0-323-99031-8.00007-7.

[27] O. Rochmawanti and F. Utaminingrum, "Chest X-Ray Image to Classify Lung diseases in Different Resolution Size using DenseNet-121 Architectures," in Proceedings of the 6th International Conference on Sustainable Information Engineering and Technology, in SIET '21. New York, NY, USA: Association for Computing Machinery, Nov. 2021, pp. 327–331. doi: 10.1145/3479645.3479667.

[28] C. Wang et al., "Pulmonary Image Classification Based on Inception-v3 Transfer Learning Model," IEEE Access, vol. 7, pp. 146533–146541, 2019, doi: 10.1109/ACCESS.2019.2946000.

[29] M. Chhabra and R. Kumar, "An Advanced VGG16 Architecture-Based Deep Learning Model to Detect Pneumonia from Medical Images," in Emergent Converging Technologies and Biomedical Systems, N. Marriwala, C. C. Tripathi, S. Jain, and S. Mathapathi, Eds., Singapore: Springer, 2022, pp. 457–471. doi: 10.1007/978-981-16-8774-7_37.

[30] M. Fontes, J. D. S. De Almeida, and A. Cunha, "Application of Example-Based Explainable Artificial Intelligence (XAI) for Analysis and Interpretation of Medical Imaging: A Systematic Review," IEEE Access, vol. 12, pp. 26419–26427, 2024, doi: 10.1109/ACCESS.2024.3367606.

[31] M. Ameen, "Explainable Mammogram Analysis with EfficientNetV2 and Grad-CAM++ for Robust Cancer Diagnosis," Diagnostics, vol. 16, no. 1, p. 105, Jan. 2026, doi: 10.3390/diagnostics16010105.