SOCIAL PRISM

Volume 3(6), 2026, 398-419
Online ISSN: 3006-8886
Print ISSN: 3006-8878

An Explainable Deep Learning Framework for Skin Cancer Detection Using Dermoscopic Images

Raees Adnan[1], Fawad Nasim[2], Muqaddas Salahuddin[3]

[1,2,3]*Faculty of Computer Science and Information Technology, Superior University, Lahore, Pakistan,*
Email: raees1995adnan@gmail.com , fawad.nasim@superior.edu.pk , muqaddassalahuddin60@gmail.com

Abstract

Skin cancer is a significant global health concern, where early and accurate detection is important for supporting effective treatment and improving patient outcomes. To perform manual examination of the skin lesions is time consuming and could rely significantly on clinical expertise, which may lead to the need of computer aided approaches helping to classify skin lesions. In this study, an automated, DL approach to classifying dermoscopic skin images as benign or malignant is proposed. In this research, a publicly available Kaggle skin cancer dataset was used, containing 3600 images with labels, half of which were malignant and the other half were benign. The data set was split into 70% training, 15% validation and 15% testing sets. The input images were preprocessed and enhanced using image processing and augmentation techniques to standardize images, enhance the diversity of training samples, and enhance model generalization. Transfer learning was explored in four different pretrained convolutional neural network (CNN) architectures. The models were trained under similar experimental conditions analysis were used for the evaluation. Of the evaluated architectures, MobileNetV2 had the best overall classification performance with an accuracy of 91.11%. The accuracy of DenseNet201 was 88.06%, higher than that of ResNet50V2 (86.81%) and Xception (85.83%). To explore the learning behavior of the models, training and validation accuracy and loss curves were also analyzed, in addition to the quantitative evaluation. Furthermore, an explainable artificial intelligence method named Grad-CAM was added to visualize the image regions that the model relied on to make its prediction, thus gaining another insight into the model's decision-making behavior of the CNN architectures. The results show that both the CNN architectures used for the experiment and the automated classification of skin lesions using them are effective, and the MobileNetV2 performs the best in the experimental framework of this study. The proposed framework can serve as a foundation for the development of efficient and interpretable computer-aided skin lesion classification systems.

Keywords: Skin Cancer Classification, Deep Learning (DL), Convolutional Neural Network (CNN), Transfer Learning, MobileNetV2, ResNet50V2, Xception, DenseNet201, Dermoscopic Images, Explainable Artificial Intelligence (XAI), Grad-CAM , Medical Image Classification.

Introduction

Skin cancer is one of the most prevalent cancer types in the world, and is a significant issue in health globally. The International Agency for Research on Cancer (IARC) estimates that in 2022, over 1.5 million new cases of skin cancer were diagnosed worldwide, with about 58,667

Received 22 June 2026; Revised 05 August 2026; Accepted 13 August 2026; Published (online) 24 August 2026

Corresponding Author Email: fawad.nasim@superior.edu.pk
DOI: https://doi.org/10.69671/socialprism.3.6.2026.202

people dying from the disease as a result. The International Agency for Research on Cancer (IARC) estimated that in 2022, there were more than 1.5 million new cases of skin cancer globally, with approximately 58,667 deaths attributed to the disease Desale & Patil, (2023). The significance of early and accurate detection of skin lesions is illustrated by these figures. It's important to remember that melanoma can grow quickly (Ahona et al., 2026), and that it's the deadliest type of skin cancer, especially because of these factors Ashraf et al., (2020). While visual examination of the lesion is important, dermatologists also use dermoscopy to examine suspicious lesions, and dermoscopy images can be challenging to interpret due to the vast variation in colour, shape, size, texture, borders, and internal structures of the lesions. These image artefacts (hairs, reflections and irregular lesion margins) may further complicate the diagnosis Jones et al., (2022).

AI and DL have emerged as promising tools to assist in identifying skin diseases (Taj et al., 2025), with a specific focus on skin lesion classification. The advancements of AI (Tauseef, Jamal, Naseer, Mahmud, et al., 2026), especially DL, have opened up novel possibilities for computer-aided skin disease classification, particularly for skin lesion classification Inthiyaz et al., (2023). While conventional machine learning algorithms (Nasim et al., 2023) require features to be carefully designed by humans to be relevant to the problem, CNNs can extract such features directly from the images(Jamal et al., 2025). The benefits of transfer learning in medical image classification lie in the fact that when there is no need for extremely large labelled medical datasets, pre-trained CNN models can effectively represent the images Arooj et al., (2022). While the potential of pretrained CNN architectures for skin lesion classification has been shown, the dataset, preprocessing, CNN architectures, and evaluation methods are different, making comparisons difficult (Tauseef, Jamal, Naseer, Akbar, et al., 2026). Although the models have good predictive performance, DL models are deemed "black-box" systems which do not offer much information regarding how each prediction is generated (Smak Gregoor et al., 2023). In medical contexts, for example, this limitation is significant as it enables clinical validation and enhances the trust placed in AI-driven systems by understanding the system's foundation for making an automated prediction.

Explainable AI (XAI) methodscan help solve this problem by offering insights into the characteristics or areas of the image that affect the model's decisions (Jamal et al., 2026). In medical image classification, visual explanations are especially useful to researchers for studying if the models are focused on clinically meaningful areas Dildar et al., (2021). One of the most common CNN prediction interpretation methods is called Grad-CAM. It generates a heatmap of image regions that help make a model's decision, using gradients from a target class (Lembhe et al., (2023). Grad-CAM can thus be applied to examine both accurate and incorrect predictions and whether the model is concentrating on the lesion or maybe maybe irrelevant areas. The classification accuracy along with the visual explanations gives a holistic perspective of the DL systems for skin lesion analysis Tahir et al., (2023).

The goal of this study is to create a DL framework capable of distinguishing between benign and malignant skin lesions from a publicly available skin lesion dataset with 3,600 images, 1,800 of which are benign and 1,800 malignant Riaz et al., (2023). The dataset is split into training, validation and test sets and preprocessing and image augmentation are performed to make images more consistent in the training set and varied in the training set. Transfer learning is applied to four pretrained CNN architectures Goyal et al., (2020). These models are different architectural designs and can be used to compare the performance of the designs under the same experimental conditions(Jamal et al., 2024). The accuracy and loss curve during training and during validation is also analysed in order to study the behaviour of the model learning. Furthermore, Grad-CAM is used to provide visual explanations for correctly and incorrectly classified images Malik et al., (2023). This approach will not only allow for the assessment of classification accuracy but also for the assessment of the regions of the images that are predicted by the model Ali et al., (2022). The primary contribution of this study is the comparative analysis of four different pretrained CNN architectures in a uniform experimental setup with Grad-CAM interpretability analysis. The results demonstrate that the MobileNetV2 model has the highest accuracy of 91.11%, followed by DenseNet201 with 88.06%, ResNet50V2 with 86.81% accuracy, and Xception with 85.83% accuracy Bansal et al., (2022). The results show that the light-weight MobileNetV2 network can achieve competitive results for the selected skin lesion classification task and can facilitate visual interpretation of the predictions Shetty et al., (2022).

The rest of this paper is structured as follows: The related work on DL and explainable artificial intelligence related to skin lesion classification is presented in section 2. Proposed methodology is detailed in Section 3, comprising of dataset preparation, dataset preprocessing, data augmentation, model architectures, training and evaluation metrics. The experimental results are given in Section 4, which contain confusion matrices, learning curves, comparison plots and Grad CAM visualisations. The conclusion of the study and future directions for research are provided in Section 5, which also points out limitations.

Literature Review

In recent years, DL and explainable (XAI) have been employed to automate the skin lesion and skin cancer classification. Such methods not only seek to enhance the diagnosis functionality, but also to make the model's prediction processes transparent (Jamal et al., 2022). The literature has explored various pretrained CNN architectures, data augmentation and explanation methods like Grad-CAM and SHAP. In recent years, the field of automated skin cancer detection has made significant progress with the use of hybrid DL, handcrafted feature integration and transformer-based architectures(Ahmad et al., 2024). Ibraheem et al. (2024) analyzed the use of machine learning to detect skin cancer and found a combination of ResNet-50, DenseNet and GoogleNet to achieve an accuracy of 97.48% on images of the ISIC database. Naeem et al. (2024) suggested SNC_Net, a hybrid system combining handcrafted and DL features for skin cancer detection. With ISIC 2019, their combination of obtained an accuracy of 97.81%, which shows the advantages of using complementary representations of features. Later studies have worked on multiclass classification and the use of sophisticated DL designs.

Ilyas et al (2025) conducted a comparison of ResNet-50, DenseNet, VGG16 and InceptionV3 for the classification of seven skin cancer types on the HAM10000 dataset with accuracy of 92.5%. In a similar study, Nawaz et al. (2025) focused on CNN-based detection of skin cancer from the images of HAM10000, and found a 96% accuracy rate with models architectures. The results clearly show the effectiveness of transfer learning and CNN-based methods in the classification of dermoscopic images with multiple classes. Additionally, Adablanu et al., (2026) investigated the fusion of CNN and Transformer model for skin cancer detection and achieved 94.1% accuracy for HAM10000 by using Vision Transformer (ViT), CNN and LSTM model. All these studies suggest that hybrid architectures and feature integration can yield a high classification accuracy,

Online ISSN: 3006-8886
Print ISSN: 3006-8878

but at the same time may also demand higher computational effort. Thus, it is still relevant to compare the performance of lightweight pre-trained CNN architectures within a uniform experimental setup, especially if one takes into account the interpretability of the networks using methods like Grad-CAM.

The skin cancer classification was performed by Mahmud et al., (2023) who tested four pre-trained DL models. To promote generalization and tackle imbalance across the classes, image augmentation was performed and to enhance the transparency of the model, the XAI techniques were integrated. The overall highest accuracy result was obtained with a combination of transfer learning and explainability, which was realized by the XceptionNet with 88.72%. The authors also called for more samples of the malignant lesions to further increase the accuracy of the models. Gamage et al. (2023) explored the classification of melanoma in the HAM10000 dataset with the Xception network. The model was then fine-tuned using bayesian hyper parameter optimization to reach the accuracy of 90.24%. The study also included Grad-CAM and Grad-CAM++ to provide visual heat maps of the image areas that were important for the model's predictions. Such explanations illustrated the effectiveness of using visualization techniques in dermatological classification using CNN. Mridha et al. (2023) proposed the CNN optimized classification system that classifies seven different types of skin cancer in the HAM10000 dataset. Various activation functions and optimization algorithms were explored, and Grad-CAM and Grad-CAM++ were applied to enhance the interpretability. The proposed approach was able to achieve an accuracy of 82%.

The study highlighted the need for more research into the quality of explanations and clinical characteristics, such as the ABCD rule, to enhance the accuracy of XAI-supported diagnostic systems. Ahmad et al. (2023) have suggested a framework with the fusion of techniques: data augmentation, transfer learning, feature fusion, optimization, and XAI. For ISIC 2018 and HAM10000, Xception and ShuffleNet were applied and Grad-CAM was used to visualize the important regions of the image. The framework achieved accuracies of 99.3% on ISIC 2018 and 91.5% on HAM10000. The findings showed good results, but the authors indicated that more computational resources would be needed and recommended further testing with other datasets and optimizations. Rehman et al. (2022) explored two tweaked versions of MobileNetV2, DenseNet201 on skin lesion classification with benign and malignant classes. Both pretrained architectures had 3 extra convolutional layers added to them. The accuracy of the modified DenseNet201 was 95.50%, with sensitivity of 93.96% and specificity of 97.03% compared with the other models pretrained.

The study showed that by optimizing the pre-designed CNN structures, the accuracy of skin lesion classification can be further improved, but the structure of the model is also getting more complex. Chowdhury et al. (2021) investigated how DL features relate to clinical diagnostic features in skin lesion classification. CNNs, spatial self-attention, and attention-based activation models were compared to models that included clinical features of ABCD. The custom CNN + CAM system achieved 82.7% accuracy on HAM10000. This study emphasized the need to consider image-level features as well as clinically meaningful features when assessing the accuracy of automated diagnostic models. Wang et al. (2022) proposed a new multimodal convolutional neural network (IM-CNN) by combining the skin lesion images, metadata, and features extracted based on the domain information. To explain the predictions based on images and metadata, SHAP and Grad-CAM were used. The proposed architecture was able to attain accuracy of

Online ISSN: 3006-8886
Print ISSN: 3006-8878

95.1% on HAM10000, showing the potentiality of using multimodal information along with explainability.

In this work the MobileNetV2, ResNet50V2, Xception and DenseNet201 architectures are compared on the same framework for preprocessing, training and testing. In addition, Grad-CAM is used to analyze the prediction of the model and the region of the image that is involved in the decision of the classification Mo et al., (2026). This quantitative performance assessment, along with the visual interpretability, has filled a gap in the existing literature, and offered a more holistic evaluation of the DL model used for benign and malignant skin lesion classification.

Proposed Methodology

The proposed DL framework for binary skin lesion classification is shown in Fig. 1. The framework is composed of six stages, namely, image preprocessing, model selection, model customization, model training, classification and performance evaluation.

- To ensure uniformity of the input data for the DL models, it is necessary to preprocess dermoscopic images at first. The images are resized to 224 x 224 x 3 pixels and are made stable for training by normalizing. In addition, data augmentations are used to enhance the image diversity and prevent overfitting. The prepared dataset is then split into train, validation and test sets.
- Four convolutional neural network (CNN) architectures are chosen for comparison. The models are based on transfer learning, where the models are pre-trained using weights learnt to recognise images in the ImageNet data set. They learn feature representations that enable the models to identify relevant visual features from dermoscopic images while alleviating the need to train a deep network from scratch.
- To use each pre-trained architecture for the binary classification task, the original classification layers are replaced by a pre-designed classification head. The head is a 2D Global Average Pooling layer, followed by a dense layer with 512 neurons and ReLU activation, a dropout layer with a rate of 0.5 and an output layer with two neurons and a Softmax activation. The output layer outputs the image label, which in this case, is either benign or malignant.
- The customized models are optimized with the Adam algorithm and learning rate of 0.0001, and a batch size of 32. The models are trained 70 times with the loss function as categorical cross-entropy loss.
- To investigate the learning behaviour and generalization of every model, performance is monitored during training, and the performance is validated. The models are then used for classification of test images that were not used in training. The evalution matrix are used to measure the performance of the four models.

Further, training and validation accuracy and loss curves are studied to evaluate model convergence and possible overfitting or underfitting of the model. The overall flow is shown in Figure 1 and the key steps of the proposed methodology are summarized.

SOCIAL PRISM
Online ISSN: 3006-8886
Print ISSN: 3006-8878

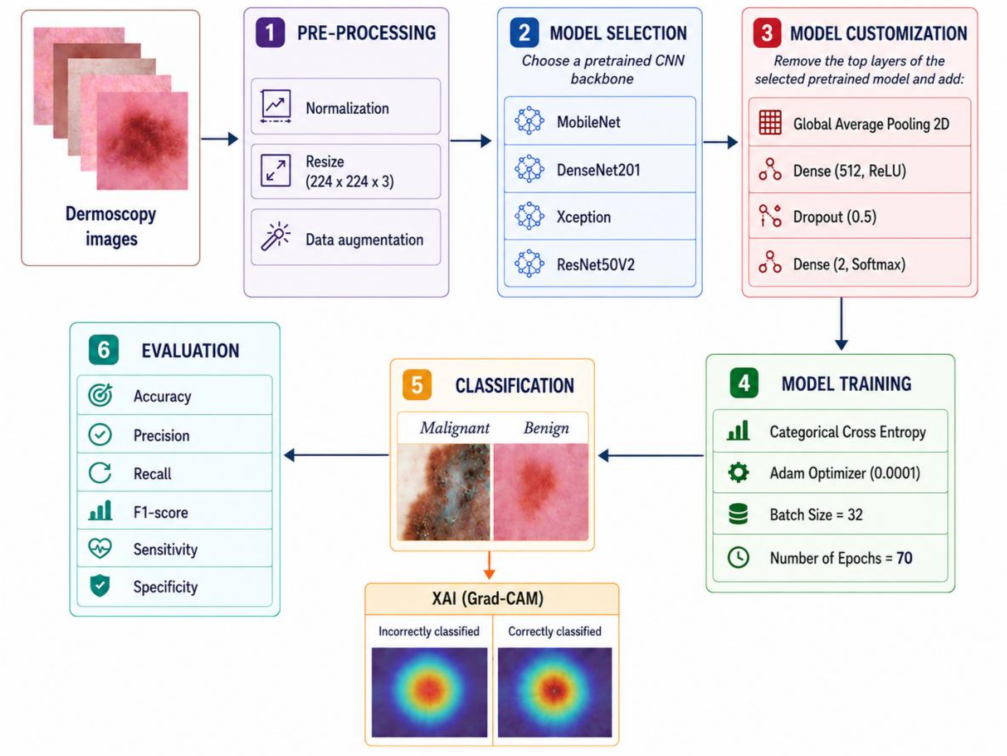

Figure 1: Proposed Method for Skin Cancer Detection

Dataset and Data Processing

Creating a suitable dataset and effective data pre-processing plays an important role in building reliable DL models for skin lesion classification. A publicly accessible skin cancer dataset was used in this study from Kaggle Fanconic. (2022). The images are dermoscopic images with two classes: Benign and Malignant. The reasons for its selection included a class distribution that was roughly equal, an appropriate number of labeled images, and its relevance to skin cancer classification studies. The dataset contains 3600 images, 1800 benign (50%) and 1800 malignant (50%) images as shown in Table 1. The distribution is balanced, which minimizes the impact of class imbalance in model training, and offers a suitable basis for binary classification.

Table 1: Brief description of dataset

Category	Images
Benign	1800
Malignant	1800
Sum	3600

All the images were prepared in the same size and format for model training, which was done by pre-processing the images. All images were resized to 224 x 224 x 3 pixels (224 x 224 x 3 colour channels) as the original images had varying sizes and appearance. Such an input size can be used with the chosen pretrained CNN architectures and the amount of computation is decreased, while maintaining relevant visual information.

Online ISSN: 3006-8886
Print ISSN: 3006-8878

The pixel values have also been normalized to ensure a consistent distribution of the data fed into the model during its training and to ensure stability in the training process. A random subset of the data is illustrated in Figure 2, which displays some example benign and malignant skin lesion images. The samples show visual differences between classes, such as lesions colour, shape, texture and border characteristics.

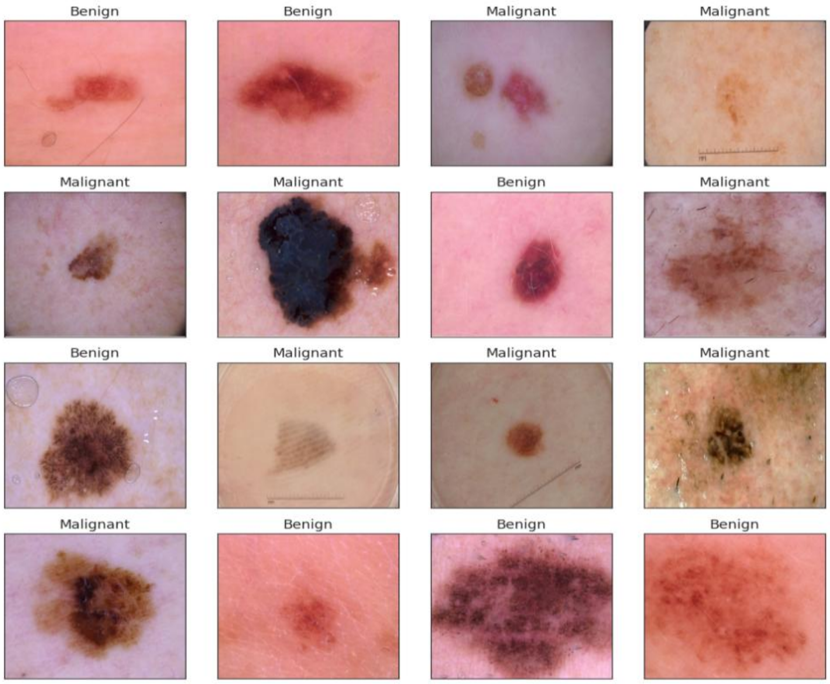

Figure 2: Random Sample of Trained Dataset

The training images were augmented with data to enhance the diversity of the data set and prevent overfitting. Even though there is equal number of benign and malignant images in the dataset, augmentation is still crucial as DL models can memorize the training samples if there is not enough diversity in the images. During the augmentation process, new images that are similar to the original images are created but the classes of the original images are not changed.

The image shifting, shearing, horizontal flipping, vertical flipping, zooming, and appropriate fill modes were the augmentation methods used in the present study. These transformations produce variations that could happen in real-world dermoscopic images and make the models learn more about features related to the lesion instead of specific orientations or locations on the image. Only the training images were augmented, validation and test images were not altered but only the necessary pre-processing steps were applied. This is to prevent that artificial images do not affect the evaluation results. Resizing, normalization and augmentation give a standardized input pipeline for the four different CNN architectures evaluated: MobileNetV2, DenseNet201, Xception and ResNet50V2. The pre-processing steps are used to enhance the generability of the model and to create a consistent experimental setting to compare the selected architectures.

Online ISSN: 3006-8886
Print ISSN: 3006-8878

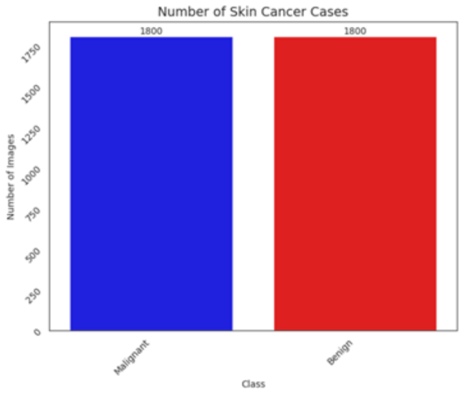

Figure 3: Number of Skin Cancer Cases

Classification

The classification stage involves using the trained DL models to differentiate between benign and malignant skin lesions. In this study, four pre-trained CNN architectures have been tested with the same experimental setup. The preprocessed and enhanced dermoscopic images were fed to the models which generated predictions for the two target classes. They were tested for their classification ability with the help of confusion matrix, training accuracy, training loss, validation accuracy and validation loss curve. These visualisations give additional information on the final predictions made by these models and the learning behaviour of the models.

MobileNetV2

MobileNetV2 is a simplified CNN network structure, which can provide efficient image classification, while having relatively low computation complexity. It adopts depthwise separable convolutions and inverted residual structures to limit the number of operations in the network while maximizing the extraction of features. The characteristics of MobileNetV2 make it suitable for applications that demand a compromise between classification accuracy and speed. In this study, the pretrained MobileNetV2 was fine-tuned for binary skin lesion classification. The model was given the resized, normalized and enhanced images, and it learned visual features of benign and malignant lesions.

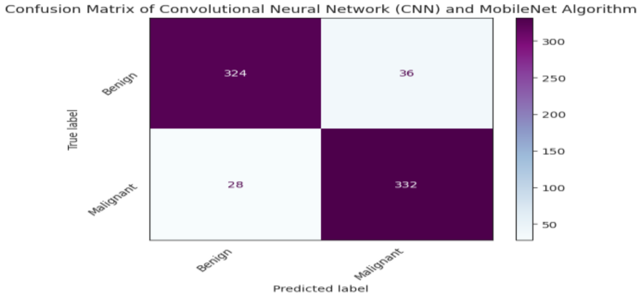

Figure 4: Confusion matrix of the MobileNetV2

The confusion matrix displays the model's predictions on independent test set. The diagonal entries are those images that are correctly classified, and the off-diagonal entries are misclassified

Online ISSN: 3006-8886
Print ISSN: 3006-8878

images. This matrix gives an overview of the model's distinguishing capabilities of benign and malignant lesions and the distribution of the classification errors that occur. The training and validation accuracy and loss curves of the MobileNetV2 model are shown in the figure below.The training and validation accuracy and loss curves of the MobileNetV2 model are presented below.

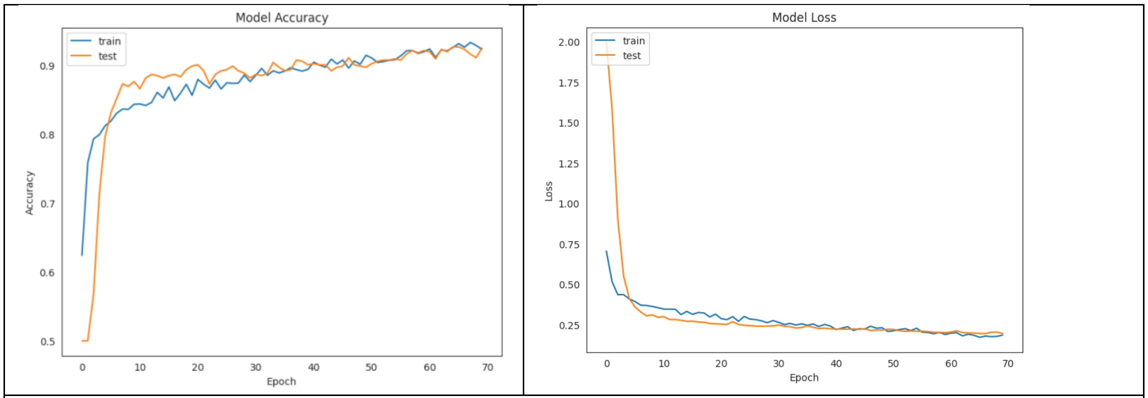

Figure 5: Accuracy and loss curves of MobileNetV2 model

The learning behaviour of MobileNetV2 as a function of the number of training epochs is shown in the training and validation curves. The accuracy curves give information about how well the classifications change and the loss curves give information about how the prediction error changes. A comparison of training and validation performance can offer an idea of model convergence, stability and generalization. The confusion matrix and learning curves are taken together to give a holistic picture of the performance of MobileNetV2.

ResNet50V2

ResNet50V2 is a deep residual CNN architecture, which incorporates the residual connections to make deep network learning easier. These connections enable the model to learn more efficiently complex hierarchical image features by allowing the passing through the network of information and gradients. A deep residual network (ResNet50V2) was chosen to test a deep residual network for the classification of skin lesions for benign and malignant skin lesions. The same preprocessing, augmentation and training procedure was used to train the model as was done for the other architectures. This provided an equal footing.

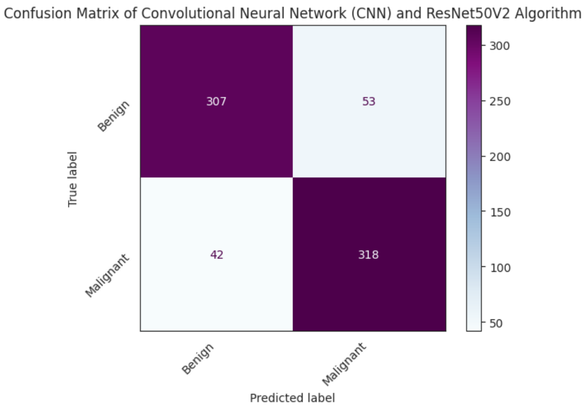

Figure 6: Confusion matrix of the ResNet50V2 model

SOCIAL PRISM
Online ISSN: 3006-8886
Print ISSN: 3006-8878

The confusion matrix is a summary of the classification result of ResNet50V2 on the test data. The diagonal cells show correct predictions, while the cells off the diagonal show incorrect predictions. The matrix is also used to study the performance of the class and to compute measures like sensitivity and specificity.

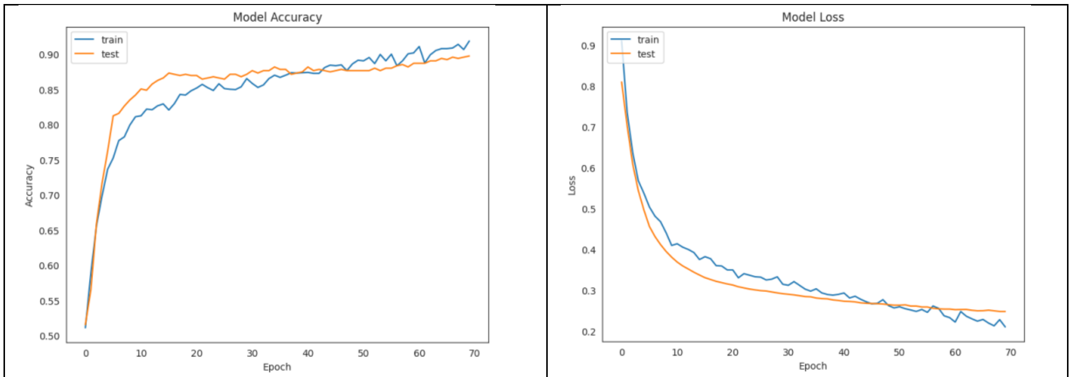

Figure 7: Accuracy and loss curves of ResNet50V2 model

The accuracy and loss curves show the performance of ResNet50V2 during the training process. Training curves show how the model learns as it trains, and validation curves give information about how well it generalizes to the validation data it never saw during training. These curves, along with the confusion matrix, give a general idea of the stability of the model's learning and classification behaviour.

Xception

Xception is a CNN architecture that uses depthwise separable convolutions. The spatial feature learning and channel-wise feature processing are decoupled in Xception, allowing it to capture fine-grained visual features with less computational redundancy than traditional convolutional layers. The same binary skin lesion classification task was used in this study with the pretrained Xception model.

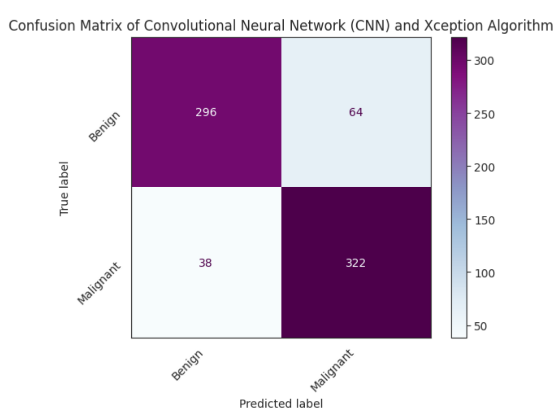

Figure 8: Confusion matrix of Xception model

The confusion matrix shows how many Xception predictions are correct and incorrect for the benign and malignant lesions. Diagonal elements are those that are correctly classified, off-diagonal elements are those that are incorrectly classified. It is a visualisation to help identify the model's performance of separating the two classes.

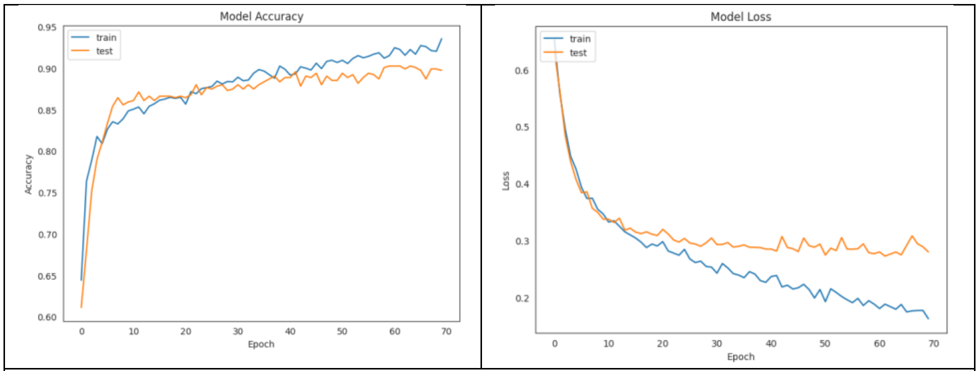

Figure 9: Accuracy and loss curves of Xception model

The curves demonstrate the learning of Xception during the model training and validation. The accuracy curves reflect the change in classification accuracy, while the loss curves reflect the change in the prediction error. These results are used to validate the accuracy of the representations learned by generalizing from the training set to the validation set. The confusion matrix and learning curve thus complement each other with regard to Xception, as the former is descriptive of the final test prediction, while the latter is descriptive of the training process.

DenseNet201

DenseNet201 is a deep CNN with dense connections between layers. All the layers pass information about features to the layers above them, which leads to re-use of features and help to propagate the information up the network. This architecture can be trained to learn detailed image representations and thus is appropriate for image classification applications with fine-grained differences in lesion classes. For pretrained DenseNet201, the fine-tuning was performed for binary classification of benign and malignant skin lesions with the same experimental procedure as the other models. The inclusion enabled comparison of the effectiveness of dense feature connectivity with the residual and depthwise-separable architectures.

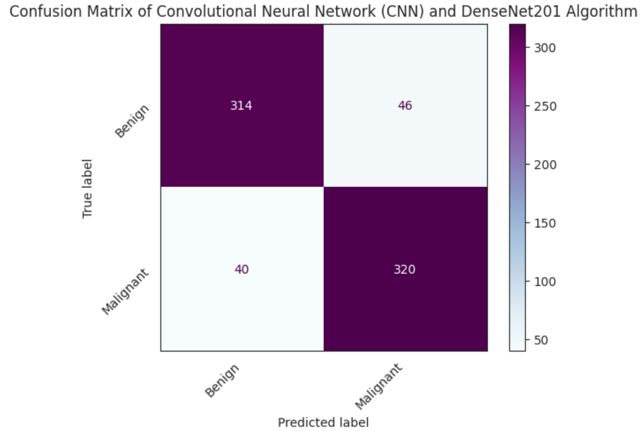

Figure 10: Confusion matrix of the DenseNet201 model

Online ISSN: 3006-8886
Print ISSN: 3006-8878

The DenseNet201 model was tested on the skin cancer dataset with a confusion matrix as shown in Figure 11.The test result of the DenseNet201 model for skin cancer classification has been shown in figure 11 as a confusion matrix. The DenseNet201 confusion matrix is a summary of the independent test set predictions by the model. Samples correctly classified are displayed as diagonal elements and samples classified wrong are displayed as off-diagonal elements. The matrix gives the idea about the capacity of model to classify the benign and malignant lesions.

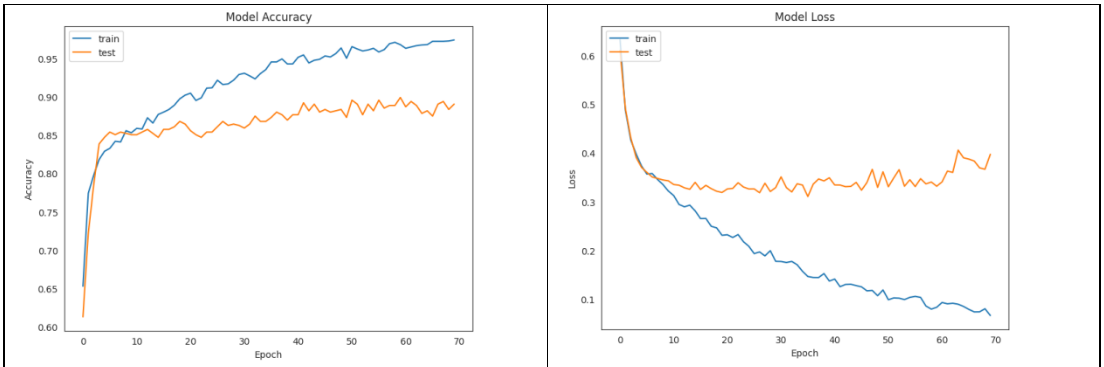

Figure 11: Accuracy and loss curves of DenseNet201 model

The following accuracy and validation loss curves show an overview of the learning behaviour of DenseNet201 during the training process. The accuracy curves represent the change in accuracy of classification, and the loss curves represent the change in prediction error. The validation performance shows an indication of the generalization ability of the model and is used in addition to the results from the test confusion matrix. Overall, the four CNN architectures provide different approaches to feature extraction and representation learning. MobileNetV2 emphasizes the computational efficiency, ResNet50V2 is based on residual learning, Xception is based on depthwise separable convolutions, and DenseNet201 is based on dense feature connectivity.

Comparative Analysis of CNN Models

To make a fair comparison for the four pretrained CNN architectures, they were tested on the same independent test set. The overall accuracy is the percentage of the correctly classified images, whereas the percentage of correct positive predictions is the precision. The recall is a measure of how well you can identify positive samples and the F1-score gives a balance of precision and recall. The sensitivity and specificity are added information that further indicates the accuracy of the model in correctly identifying the respective classes.

Table 1: Comparative Performance of the CNN Models

CNN Model	Accuracy	Precision	Recall	F1-Score	Sensitivity	Specificity
MobileNetV2	91.11%	91.13%	91.11%	91.11%	90.00%	92.22%
ResNet50V2	86.81%	87.00%	87.00%	87.00%	85.27%	88.33%
DenseNet201	88.06%	88.00%	88.00%	88.00%	87.22%	88.88%
Xception	85.83%	86.00%	86.00%	86.00%	82.22%	89.44%

The accuracy 91.11% respectively for MobileNetV2. It also obtained a sensitivity of 90.00% and a specificity of 92.22%, which is relatively well-balanced between both classes. The Dense-

Online ISSN: 3006-8886
Print ISSN: 3006-8878

Net201 model had a second-highest accuracy of 88.06%, and precision, recall, and F1-score of around 88.00%. The sensitivity and specificity of its performance were 87.22% and 88.88%, respectively. The accuracy of ResNet50V2 is 86.81%. The sensitivity was 85.27% and the specificity was 88.33%. Xception had the lowest overall accuracy, which was 85.83%. It had a relatively good specificity value of 89.44%, but the lowest sensitivity value of 82.22% among the models evaluated. This means that the classification setting in this study results in relatively poor performance for identifying the positive class.

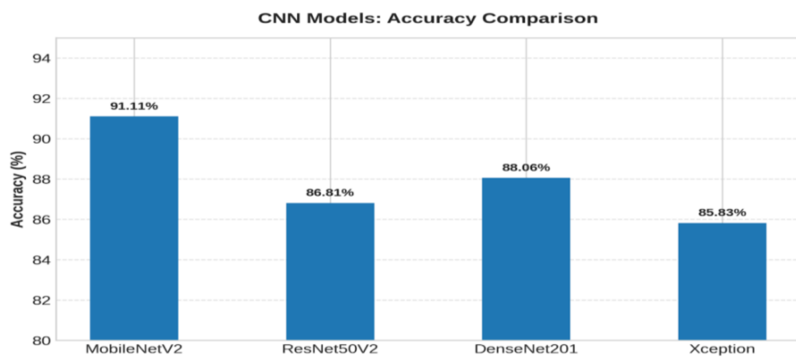

Figure 12: Graph of CNN models across accuracy

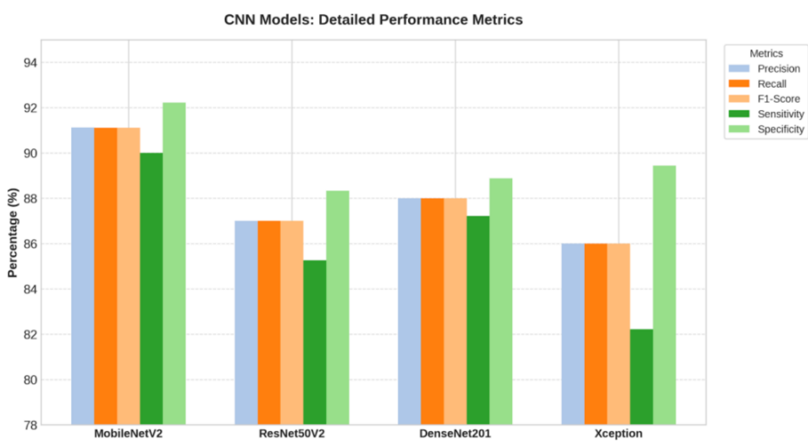

Figure 13: Graph of the CNN models

The compared metrics are graphically presented in Figures 12 and 13. The results show that the overall performance is better for MobileNetV2, followed by DenseNet201, ResNet50V2 and Xception. The results indicate that the MobileNetV2 lightweight architecture is very efficient for the chosen problem of skin lesion classification.

Grad-CAM-Based Model Interpretability

A quantitative performance measure reflects the accuracy of the model but does not explain which parts of the image are important to the model's predictions. Therefore, to gain visual interpretation of the prediction made by the four CNN models, the Grad-CAM method was used. Grad-CAM generates class-specific heatmaps using the gradients of the predicted class with respect to the feature maps of a convolutional layer. The heatmaps that are produced give an idea

Online ISSN: 3006-8886
Print ISSN: 3006-8878

of regions that are more effective in the model's prediction. This enables the spatial focus of attention of the model to be analyzed in the context of skin lesion classification, and also offers an indication of whether predictions relate to the skin lesion itself. For both correct and incorrect classification, grad-CAM visualisations were created. If the models make correct predictions, it is possible to consider whether the models consider the relevant lesion characteristics; if they make incorrect predictions, it is possible to consider whether the models consider ambiguous characteristics or less informative parts.

MobileNetV2 Grad-CAM Analysis

The visualisations of Grad-CAM are shown for representative correctly and incorrectly classified images of the MobileNetV2 model in Figure 14. The heatmaps give a representation of the areas that affect the model's predictions. If the examples are correctly classified then the model can be analysed to see if the model is focused around relevant parts of the lesion. Examples that are not classified correctly add to the knowledge about possible classification uncertainty sources. If the highlighted regions are outside of the lesion or are focused on the ambiguous regions, it can be a sign that the learned features were not enough to separate the two classes.

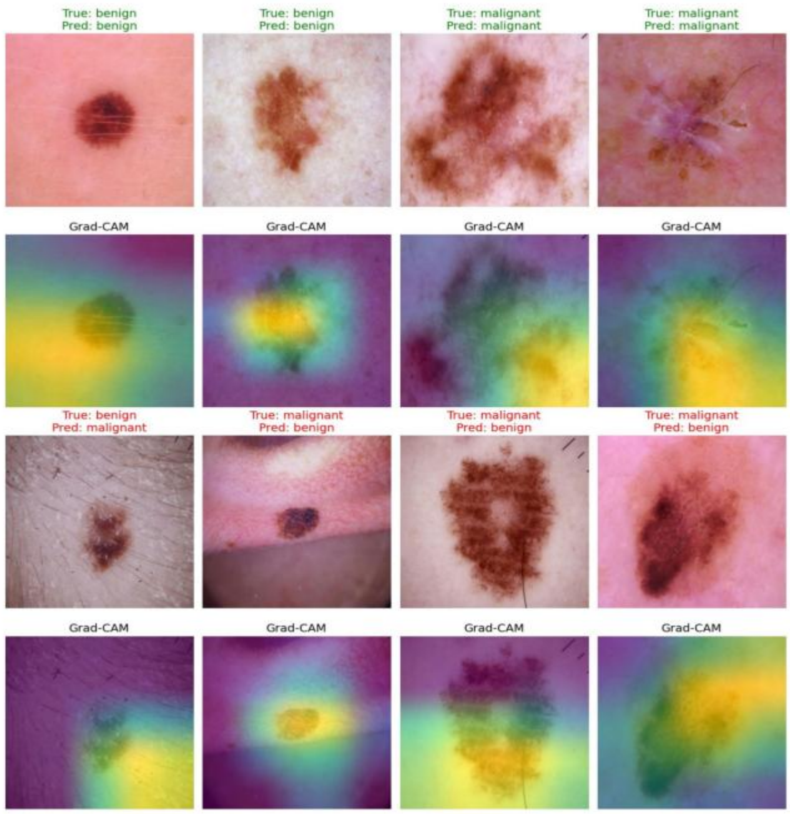

Figure 14: Grad-CAM of MobileNetV2 model

Correctly and incorrectly classified instances from the MobileNetV2 model with Grad-CAM visualisation are presented in figure 14. Grad-CAM results give an image-level view of the classification behaviour of MobileNetV2, while the numerical performance of the model is displayed on the right.

Online ISSN: 3006-8886
Print ISSN: 3006-8878

ResNet50V2 Grad-CAM Analysis

In Figure 16, representative Grad-CAM visualisations produced by ResNet50V2 are illustrated. Images that are correctly classified can be viewed to determine if the remaining architecture is geared toward relevant characteristics of the lesions. The misclassified images help to understand what parts of the image the model might have focused on but not on the most discriminative parts.

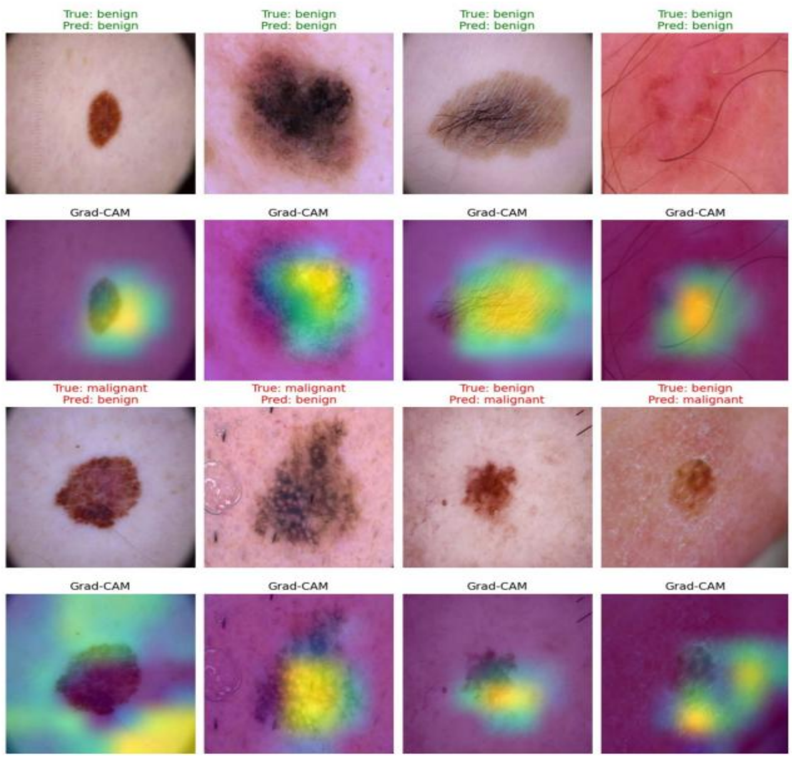

Figure 15: Grad-CAM of ResNet50V2 model

The visualisation of the Grad-CAM for correctly and incorrectly classified instances for the ResNet50V2 model are shown in Figure 15. The visual explanations give extra context to the confusion matrix and quantitative explanations to the spatial focus of the model.

Xception Grad-CAM Analysis

In Figure 17, the Grad-CAM results of Xception are presented. The visualisations enable the regions that were classified correctly or incorrectly based on the model to be looked at. Accurate predictions can be used to indicate attention to relevant lesion characteristics and inaccurate predictions can be used to indicate potentially problematic regions in the image.

SOCIAL PRISM
Online ISSN: 3006-8886
Print ISSN: 3006-8878

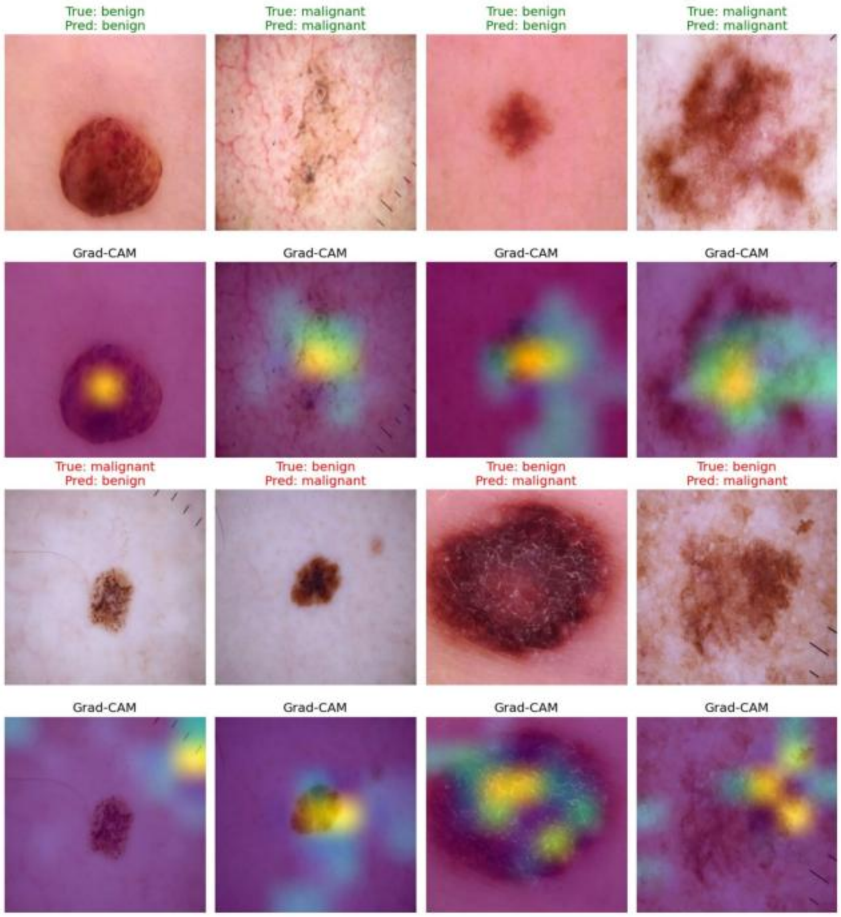

Figure 16: Grad-CAM of Xception model

Grad-CAM visualisation of correctly and incorrectly classified instances by the Xception model (figure 16) The visual explanations are not only a supplement to the numerical evaluation, but also give information about the spatial regions, in which the classification decisions of Xception are made.

DenseNet201 Grad-CAM Analysis

Figure 18 shows visualisations of DenseNet201 using Grad-CAM. The heatmaps give an idea of how the highly connected architecture is concentrating its predictions on relevant lesion regions. Misclassified examples can also indicate unclear patterns or regions that can lead to classification mistakes.

All four CNN architectures can be interpreted using the Grad-CAM. In contrast, with conventional metrics, predictive performance is measured, while Grad-CAM also supplies visual information regarding the regions related to the model's decisions. The analysis of both correct and incorrect predictions is thus informative for a better understanding of model behaviour.

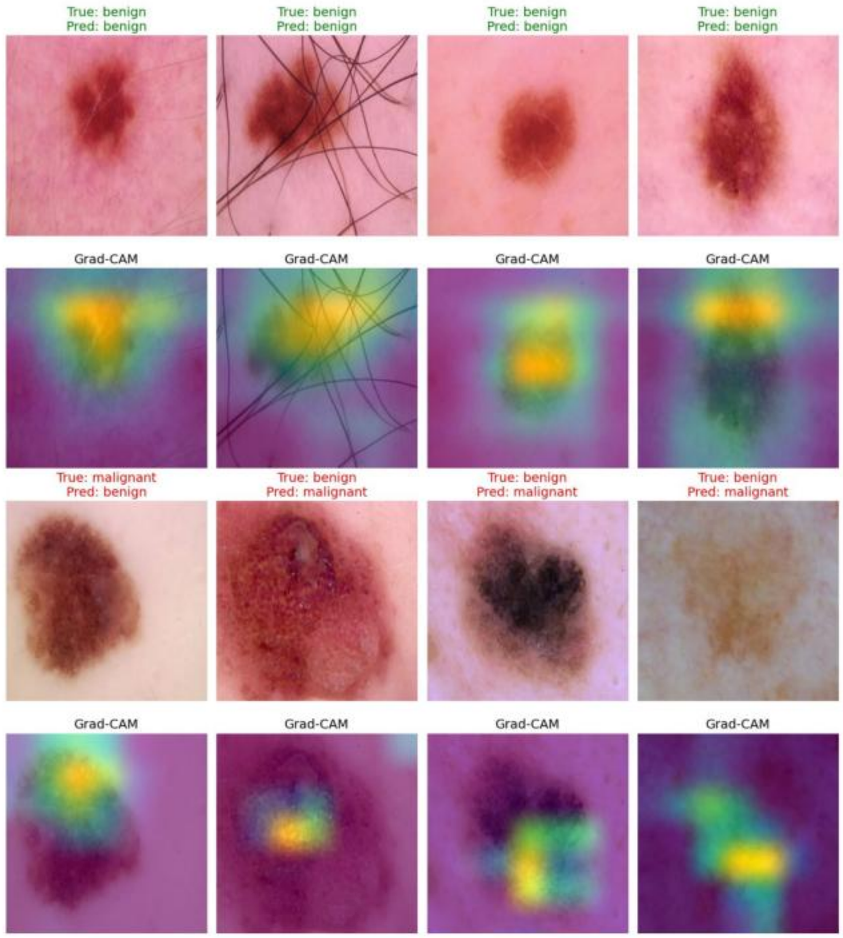

Figure 17: Grad-CAM of DenseNet201 model

Discussion

The experimental results showed that all four CNN architectures were able to differentiate between benign and malignant skin lesions but with significant variations in accuracy. MobileNetV2 was the best performer with an accuracy of 91.11% and an F1 score of 91.11%. The precision, sensitivity, and specificity are relatively well balanced, with a value of 91.13%, 90.00% and 92.22%, respectively. The great performance of MobileNetV2 is remarkable as it is a relatively lightweight architecture. The efficient convolutional architecture allows meaningful feature extraction in a way that is computationally more simple than many deeper CNN architectures. The findings therefore indicated that the more complex and deep a model is, the more does not mean the more better will the model perform. Next, DenseNet201 had the second highest accuracy of 88.06%. It is highly connected and allows for feature re-use and propagation of information through the network, potentially benefiting the learning of subtle visual features. Under the experimental conditions of this study, however, it did not perform as well as MobileNetV2. ResNet50V2 was able to achieve 86.81% accuracy, about 87.00% precision, recall, and F1-score. While the residual network helps deep networks train, the results reveal that this method was not superior to the MobileNetV2 or DenseNet201 for the classification task being used here. The lowest overall accuracy was obtained by Xception (85.83%). Although its specificity of 89.44%

was quite high, the sensitivity was the lowest at 82.22% among models evaluated. This difference indicates that Xception had only moderately good accuracy for the other class while it was not as good for the positive class.

The confusion matrices reinforce the overall ranking of performances. The model of “MobileNetV2” made the most correct classifications while making a relatively small number of errors compared to the other models. DenseNet201 also resulted in relatively equal distribution of correct predictions while ResNet50V2 and Xception gave rise to more classification errors. The confusion matrices then give the level of information at the class level that complements the overall performance metrics. The training and validation accuracy and loss curves give further insights into model learning behaviour. These curves can be used to analyze the development of model performance over the course of the training epochs, and to gauge the model's rate of convergence and its level of generalization. It is crucial to assess learning curves in conjunction with the independent test results because high training performance does not necessarily mean high performance on unseen images. Multiple evaluation metrics are especially crucial for medical image classification. While accuracy measures the overall correctness, precision, recall and F1-score offer a more detailed view of the quality of the predictions. Sensitivity and specificity are additional terms used to describe the performance of the model on the classes and to highlight the differences in the model's performance on the two classes.

The interpretability evaluation is added by introducing the Grad-CAM analysis in the experimental evaluation. Grad-CAM does not only give numeric data on the regions relevant to the model predictions but also visual data. Correctly classed examples may help determine if models are targeting important characteristics of the lesion correctly, and incorrectly classed examples can help determine if there are any ambiguous or irrelevant regions that correspond to prediction errors. Therefore, the evaluation of the four CNN architectures is more comprehensive combining the quantitative evaluation and visual interpretation. Overall, MobileNetV2 offered a good balance of classification performance out of the evaluated architectures. The results also confirm that there is no superior convergence between a deeper or more complex model, and the better performance in every medical image classification problem.

Limitations and Future Work

The results of this study should be interpreted with a few caveats. First, the study is based on a single publicly available dataset, potentially missing the diversity of skin lesions seen in other imaging devices, in other populations or in other clinical settings. The generalizability of the models needs to be further investigated with independent and multi-centre datasets. Second, data augmentation was performed to improve generalization performance, but variations in image quality, image characteristics of lesions, and acquisition conditions can still impact model performance. Larger and more diverse data sets, as well as external validation data sets, should be used to test the models in future studies. Third, various CNN architectures are different in terms of the number of computations they require.

While MobileNetV2 is efficient, there is potential for further compression, pruning and/or quantization to enable deployment to resource-constrained devices. Last, Grad-CAM offers helpful visual explanations that should not be interpreted as a precise segmentation of the lesions or as concrete clinical reasoning. Other XAI approaches such as Grad-CAM++, SHAP, Integrated

Gradients or quantitative assessment of explanation quality could be explored in future research. Combining complementary models with ensemble and hybrid architectures might be explored to assess the classification robustness.

Clinical Implications

It is proposed that the framework can be used as a computer-aided tool for the assessment of skin lesions. By implementing an automatic classification system, the system could help healthcare personnel to make quick preliminary predictions and aid in the evaluation of lesions of concern. The proposed models are not a substitute for a dermatologist or a clinical diagnosis, however. Before clinical use, a large-scale external validation, prospective testing and assessment in a broad variety of patients and imaging settings would be needed to validate this for clinical use. There are also regulatory and clinical requirements that need to be taken into account. The MobileNetV2 architecture is relatively light-weight which can be beneficial for future deployment in environments with limited computational resources. Overall, the use of Grad-CAM can enhance transparency by enabling visualizing the image regions related to the model's predictions, which can be useful for clinicians and researchers. However, use of visual explanations should be thought of as a supplementary piece of evidence and not as a definitive clinical reason.

Conclusion

In this study they introduced a DL solution for binary classification of skin lesions into benign and malignant with four pretrained CNN architectures: MobileNetV2, ResNet50V2, Xception, and DenseNet201. All models were tested under equal preprocessing, augmentation, training and testing with accuracy, precision, recall, F1 score, sensitivity and specificity. MobileNetV2 showed the best performance across all architectures, with an accuracy of 91.11%, while DenseNet201, ResNet50V2, and Xception were 88.06%, 86.81%, and 85.83% accurate, respectively. The MobileNetV2 model also performed best with the highest overall F1-score and showed a relatively good sensitivity and specificity. The results obtained show that a light CNN architecture is able to achieve high performance on the chosen classification task of skin lesions without depending on the greater depth or complexity of more complex CNN architectures. Grad-CAM was employed to give the visual explanations of the model predictions along with the quantitative evaluation. This solution allowed the analysis of the image segments corresponding to the correct and incorrect classifications, giving more insights into the model behaviour and performance as shown in the heatmaps. Predictive performance and visual interpretability give a more thorough assessment of the DL methods used for skin lesion classification. Further research using larger and more diverse datasets, model generalization, more alternative explainability methods, and efficient or ensemble architectures should be explored. Such systems will need to be clinically validated under real-world conditions, before they can be recommended for practical use in diagnostic support. In conclusion, the results demonstrate that Pytorch-based CNNs, specifically MobileNetV2, have great potential to build effective and interpretable computer-aided skin lesion classification systems.

References

[1]. Desale RP, Patil PS. An automated hybrid attention based deep convolutional capsule with weighted autoencoder approach for skin cancer classification. The Imaging Science Journal. 2023 Jun 29:1–5.

[2]. Mahmud, F., Mahfiz, M.M., Kabir, M.Z., & Abdullah, Y. (2023). An Interpretable Deep Learning Approach for Skin Cancer Categorization. ArXiv, abs/2312.10696.

[3]. Ahona, M. H., Jamal, A., Tauseef, F., Mahmud, M., & Indika, M. (2026). AI-Driven Discovery of Novel Biomarkers for Early Cardiovascular Disease Detection. Well Testing, 35, 04.

[4]. Ashraf R, Kiran I, Mahmood T, Butt AU, Razzaq N, Farooq Z. An efficient technique for skin cancer classification using deep learning. In2020 IEEE 23rd International Multitopic Conference (INMIC) 2020 Nov 5 (pp. 1–5). IEEE.

[5]. Taj, F., Muqaddas, R., Yousaf, A., Nasim, M. F., Naeem, M., & Nawaz, E. (2025, October). Ensembled Deep Learning (DL) Methods for Detecting Skin Cancer Using CNN Algorithm with Multi Model. In 2025 International Conference on Engineering and Emerging Technologies (ICEET) (pp. 1-6). IEEE.

[6]. Jones OT, Matin RN, van der Schaar M, Bhayankaram KP, Ranmuthu CK, Islam MS, et al. Artificial intelligence and machine learning algorithms for early detection of skin cancer in community and primary care settings: a systematic review. The Lancet Digital Health. 2022 Jun 1; 4(6):e466–76. https:// doi.org/10.1016/S2589-7500(22)00023-1 PMID: 35623799

[7]. Tauseef, F., Jamal, A., Naseer, F., Mahmud, M., Mahmood, M. M., Islam, M. R., & Chowdhury, S. A. (2026). MACHINE LEARNING MODELS FOR PREDICTING PATIENT OUTCOMES IN INTENSIVE CARE UNITS: A CASE STUDY IN US HOSPITALS. Cuestiones de Fisioterapia, 55(1), 88-105.

[8]. Inthiyaz S, Altahan BR, Ahammad SH, Rajesh V, Kalangi RR, Smirani LK, et al. Skin disease detection using deep learning. Advances in Engineering Software. 2023 Jan 1; 175:103361.

[9]. Jamal, A., Tauseef, F., Naseer, F., Akbar, A., Jabeen, M., & Nasim, F. (2025). A convolutional neural network approach for diabetic retinopathy detection from retinal images: A critical review of deep learning techniques in ophthalmic diagnosis. Multidisciplinary Surgical Research Annals, 3(4), 3100–3128. https://doi.org/10.5281/zenodo.20587510

[10]. Arooj S, Khan MF, Khan MA, Khan MS, Taleb N. Machine Learning Models for the Classification of Skin Cancer. In2022 International Conference on Business Analytics for Technology and Security (ICBATS) 2022 Feb 16 (pp. 1–8). IEEE.

[11]. Tauseef, F., Jamal, A., Naseer, F., Akbar, A., Jabeen, M., & Nasim, M. F. (2026). Predictive AI healthcare analytics for early disease detection: Leveraging AI and machine learning models to identify high-risk patients and improve preventive healthcare strategies. Annual Methodological Archive Research Review.

[12]. Smak Gregoor AM, Sangers TE, Bakker LJ, Hollestein L, Uyl–de Groot CA, et al. An artificial intelligence based app for skin cancer detection evaluated in a population based setting. NPJ digital medicine. 2023 May 20; 6(1):90.

[13]. Jamal, A., Mahmud, M., Tauseef, F., Sozib, H. M., Shafi, S. B., & Tariquzzaman, M. (2026, April). Explainable AI for Predicting Patient Readmission in Hospitals. In *2026*

9th International Conference on Inventive Computation Technologies (ICICT) (pp. 1961-1967). IEEE.
[14]. Han H, Wang WY, Mao BH. Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning. InInternational conference on intelligent computing 2005 Aug 23 (pp. 878–887). Berlin, Heidelberg: Springer Berlin Heidelberg.
[15]. Gamage, U. Isuranga, S. De Silva and D. Meedeniya, “Melanoma Skin Cancer Classification with Explainability,” 2023 3rd International Conference on Advanced Research in Computing (ICARC), Belihuloya, Sri Lanka, 2023, pp. 30–35, https://doi.org/10.1109/ICARC57651.2023.10145622.
[16]. Wang S, Yin Y, Wang D, Wang Y, Jin Y (2022) Interpretabilitybased multimodal convolutional neural networks for skin lesion diagnosis. IEEE Trans Cybernet 52(12):12623–12637. https:// doi.org/10.1109/TCYB.2021.3069920.
[17]. Dildar M, Akram S, Irfan M, Khan HU, Ramzan M, Mahmood AR, et al. Skin cancer detection: a review using deep learning techniques. International journal of environmental research and public health. 2021 May 20; 18(10):5479.
[18]. Lembhe A, Motarwar P, Patil R, Elias S. Enhancement in Skin Cancer Detection using Image Super Resolution and Convolutional Neural Network. Procedia Computer Science. 2023 Jan 1; 218:164–73.
[19]. Tahir M, Naeem A, Malik H, Tanveer J, Naqvi RA, Lee SW. DSCC_Net: Multi-Classification Deep Learning Models for Diagnosing of Skin Cancer Using Dermoscopic Images. Cancers. 2023 Apr 6; 15 (7):2179.
[20]. Chowdhury T, Bajwa ARS, Chakraborti T, Rittscher J, Pal U (2021) Exploring the Correlation Between Deep Learned and Clinical Features in Melanoma Detection. In: Papież BW, Yaqub M, Jiao J, Namburete AIL, Noble JA (eds) Medical Image Understanding and Analysis MIUA 2021 Lecture Notes in Computer Science. Springer, Cham, p 12722.
[21]. Riaz S, Naeem A, Malik H, Naqvi RA, Loh WK. Federated and Transfer Learning Methods for the Classification of Melanoma and Nonmelanoma Skin Cancers: A Prospective Study. Sensors. 2023 Oct 13; 23(20):8457. https://doi.org/10.3390/s23208457 PMID: 37896548
[22]. Goyal M, Knackstedt T, Yan S, Hassanpour S. Artificial intelligence-based image classification methods for diagnosis of skin cancer: Challenges and opportunities. Computers in biology and medicine. 2020 Dec 1; 127:104065.
[23]. Jamal, A., Tauseef, F., Naseer, F., & Ahmad, J. (2024). Business analytics for healthcare cost optimization: A data-driven study on improving financial sustainability in healthcare systems. *Spectrum of Engineering Sciences*, 723–736.
[24]. Malik H, Anees T, Din M, Naeem A. CDC_Net: Multi-classification convolutional neural network model for detection of COVID-19, pneumothorax, pneumonia, lung Cancer, and tuberculosis using chest Xrays. Multimedia Tools and Applications. 2023 Apr; 82(9):13855–80. https://doi.org/10.1007/s11042- 022-13843-7 PMID: 36157356
[25]. Zia Ur Rehman M, Ahmed F, Alsuhibany SA, Jamal SS, Zulfiqar Ali M, Ahmad J (2022) Classification of Skin cancer lesions using explainable deep learning. Sensors 22:6915. https://doi.org/10. 3390/s22186915.
[26]. Ali K, Shaikh ZA, Khan AA, Laghari AA. Multiclass skin cancer classification using EfficientNets–a first step towards preventing skin cancer. Neuroscience Informatics. 2022 Dec 1; 2(4):100034.

[27]. Bansal P, Garg R, Soni P. Detection of melanoma in dermoscopic images by integrating features extracted using handcrafted and deep learning models. Computers & Industrial Engineering. 2022 Jun 1; 168:108060.
[28]. Shetty B, Fernandes R, Rodrigues AP, Chengoden R, Bhattacharya S, Lakshmanna K. Skin lesion classification of dermoscopic images using machine learning and convolutional neural network. Scientific Reports. 2022 Oct 28; 12(1):18134.
[29]. Ahmad N, Shah JH, Khan MA, Baili J, Ansari GJ, Tariq U, Kim YJ, Cha J-H (2023) A novel framework of multiclass skin lesion recognition from dermoscopic images using deep learning and explainable AI. Front Oncol 13:1151257. https://doi.org/10.3389/fonc.2023.1151257.
[30]. Fanconic. (2022). Skin cancer: Malignant vs. benign [Dataset]. Kaggle. https://www.kaggle.com/datasets/fanconic/skin-cancer-malignant-vs-benign
[31]. Dahou A, Aseeri AO, Mabrouk A, Ibrahim RA, Al-Betar MA, Elaziz MA. Optimal Skin Cancer Detection Model Using Transfer Learning and Dynamic-Opposite Hunger Games Search. Diagnostics. 2023 Apr 28; 13(9):1579. https://doi.org/10.3390/diagnostics13091579 PMID: 37174970
[32]. Mridha K, Uddin MM, Shin J, Khadka S, Mridha MF. An Interpretable Skin Cancer Classification Using Optimized Convolutional Neural Network for a Smart Healthcare System. IEEE Access. 2023 Apr 24.
[33]. Jamal, A., Akbar, Z., Akbar, S., Niaz, S., & Tauseef, F. (2022). Predicting human-GenAI collaboration effectiveness: A machine learning investigation of skill configurations, trust, and work design. *Migration Letters*, *19*(S8), 2303–2324.
[34]. Ahmad, J., Tauseef, F., & Akbar, Z. (2024). Predictive analytics for AI-assisted patient no-show management and clinic revenue optimization: A simulation-based research. *Migration Letters*, *21*(S13), 1901–1924.
[35]. Ibraheem, I., Ramay, S. A., Abbas, T., Hassan, R. ul, & Khan, S. (2024). Identification of skin cancer using machine learning. Computers, Materials & Continua, 82(2), 2473–2495.
[36]. Naeem, A., Anees, T., Khalil, M., Zahra, K., Naqvi, R. A., & Lee, S.-W. (2024). SNC_Net: Skin cancer detection by integrating handcrafted and deep learning-based features using dermoscopy images. Mathematics, 12(7), Article 1030.
[37]. Ilyas, U., Fuzail, M., Abid, M. K., Khan, T. F., Naeem, A., & Aslam, N. (2025). Multiclass skin cancer classification using ResNet and dermoscopic imaging. Spectrum of Engineering Sciences, 3(6), 671–685.
[38]. Nawaz, K., Zanib, A., Shabir, I., Li, J., Wang, Y., Mahmood, T., & Rehman, A. (2025). Skin cancer detection using dermoscopic images with convolutional neural network. Scientific Reports, 15, Article 7252.
[39]. Mo, T., Zins, M., Goldberg, M., Ribet, C., Kab, S., Schreiver, I., Siewert, K., Ezzedine, K., Schüz, J., & Foerster, M. (2026). Tattoos and risk of cutaneous melanoma and non-melanoma skin cancer in France. JNCI: Journal of the National Cancer Institute, 118(3), 476–484. https://doi.org/10.1093/jnci/djaf332
[40]. Adablanu, S., Barman, U., Das, D., & Dweh, T. J. (2026). Transforming skin cancer detection with AI-based convolutional and transformer models. iRADIOLOGY, 4, 51–62. https://doi.org/10.1002/ird3.70058