



Article

SCCM: An Interpretable Enhanced Transfer Learning Model for Improved Skin Cancer Classification

Md. Rifat Aknda ¹, Fahmid Al Farid ², Jia Uddin ³, Sarina Mansor ^{2,*} and Muhammad Golam Kibria ^{1,*}

¹ Department of Computer Science and Engineering, University of Liberal Arts Bangladesh (ULAB), Dhaka 1207, Bangladesh; rifat.aknda.cse@ulab.edu.bd

² Centre for Image and Vision Computing (CIVC), COE for Artificial Intelligence, Faculty of Artificial Intelligence and Engineering (FAIE), Multimedia University, Cyberjaya 63100, Selangor, Malaysia; fahmid.farid@mmu.edu.my

³ AI and Big Data Department, Woosong University, Daejeon 34606, Republic of Korea; jia.uddin@wsu.ac.kr

* Correspondence: sarina.mansor@mmu.edu.my (S.M.); golam.kibria@ulab.edu.bd (M.G.K.)

Abstract

Skin cancer is the most common cancer worldwide, for which early detection is crucial to improve survival rates. Visual inspection and biopsies have limitations, including being error-prone, costly, and time-consuming. Although several deep learning models have been developed, they demonstrate significant limitations. An interpretable and improved transfer learning model for binary skin cancer classification is proposed in this research, which uses the last convolutional block of VGG16 as the feature extractor. The methodology focuses on addressing the existing limitations in skin cancer classification, to support dermatologists and potentially saving lives through advanced, reliable, and accessible AI-driven diagnostic tools. Explainable AI is incorporated for the visualization and explanation of classifications. Multiple optimization techniques are applied to avoid overfitting, ensure stable training, and enhance the classification accuracy of dermoscopic images into benign and malignant classes. The proposed model shows 90.91% classification accuracy, which is better than state-of-the-art models and established approaches in skin cancer classification. An interactive desktop application integrating the model is developed, enabling real-time preliminary screening with offline access.

Keywords: good health and well-being; skin cancer; transfer learning; explainable AI; dermoscopic image analysis



Academic Editor: Hiroyuki Yoshida

Received: 23 June 2025

Revised: 17 July 2025

Accepted: 31 July 2025

Published: 5 August 2025

Citation: Aknda, M.R.; Farid, F.A.; Uddin, J.; Mansor, S.; Kibria, M.G. SCCM: An Interpretable Enhanced Transfer Learning Model for Improved Skin Cancer Classification. *BioMedInformatics* **2025**, *5*, 43. <https://doi.org/10.3390/biomedinformatics5030043>

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Skin cancer stands as the most common form of cancer worldwide and prevalence shows a consistent increase due to causes such as recurring exposure to ultraviolet radiation and genetic reasons [1]. Malignant forms of cancer, and particularly melanoma, are responsible for most mortality due to skin cancer [2]. With over 1.5 million skin cancer cases identified globally in 2020, and more than 120,000 associated deaths documented, the disease presents a significant health challenge [3]. Early diagnosis holds good patient outcomes, as mortality increases drastically for later stages [4]. Conventional diagnosis methods, visual assessment, and biopsies are subjective, time-consuming, prone to human error, and could result in misdiagnosis [5,6]. Therefore, the need has been growing in applying artificial intelligence and deep learning methods in the preliminary screening of skin cancer diagnosis. The existing deep learning-based approaches for skin cancer classification have numerous drawbacks. These include sub-optimal classification performance,

overfitting, failure to generalize well, lack of result interpretation, being computationally expensive, and the unavailability of real-time deployable models, which further restrict their practical uses [1,7]. So, the need remains for an accurate, interpretable, and clinically deployable model that can perform efficient, reliable, and real-time classification, especially discriminating malignant and benign skin cancer stages.

To overcome these constraints, this research introduced an interpretable and enhanced transfer learning model for binary skin cancer classification. Pre-trained VGG16 convolutional neural network architecture used as the feature extractor and for transfer learning purposes combined with newly introduced customized dense layers, state-of-art activation functions, and various regularization strategies for enhancing classification accuracy. Explainable AI was incorporated for the visualization and explanation of each classification. The proposed model is deployed through a web-based system for real-time clinical deployment, which allows for dermoscopic image upload for real-time analysis and reliable classification. The methodology follows a structured process to achieve optimal performance of the model. Labeled data was acquired from the online Kaggle repository [8]. All except for last convolutional block of VGG16 were frozen in order to keep learning intact while enabling task-adaptive fine-tuning. The classifier head was designed with two dense layers using LeakyReLU activation for preventing a vanishing gradient problem. Batch normalization, dropout, and kernel regularization were applied for improving generalizability and reducing overfitting. The model was trained on the acquired dataset and the performance of the model was evaluated utilizing various standard metrics. The model was integrated in a desktop app for offline access and real-time classification.

The proposed model utilizes the pre-trained VGG16 deep learning model as a feature extractor with transfer learning. VGG16 was selected due to its robust deep architecture, which utilizes small convolutional filters, which is highly effective at extracting complex patterns from dermoscopic images, coupled with its consistent performance in diverse medical image analysis applications. The transfer learning methodology involves leveraging pre-trained knowledge and fine-tuned the final convolutional block of the model for specific tasks of skin cancer classification to improve performance even with limited medical data. Augmented Grad-CAM++ integration, an advanced explainable AI approach that provides interpretability for these predictions, promotes trust, confidence, and reliability in both clinical diagnostic processes and the automated systems.

2. Literature Review

The latest advancements in binary skin cancer classification, largely utilizing deep learning and computer vision, have seen transfer learning emerge as a pivotal strategy. Researchers have frequently utilized and fine-tuned established convolutional neural network architectures such as VGG16, VGG19, InceptionNet, ResNet, and other architectures.

A modified VGG16 model with added layers, used Kaggle dataset and expanded training set to 5636 images with augmentation, attained 89.09% accuracy, surpassing the base VGG16 [1]. Another study utilized the same dataset, fine-tuned VGG16, reached 84.24% accuracy by using texture and shape features [9]. Transfer learning with CNNs on the same dataset reached 86.65% accuracy, demonstrating CNN potentials [10]. Authors in a study fine-tuned the InceptionNet model, reporting 85.94% accuracy using the same dataset, indicating its suitability for decision support systems [11]. Pre-trained deep neural networks, such as AlexNet, ResNet-18, SqueezeNet, and ShuffleNet were used for skin cancer binary classification on the same dataset, and reached an accuracy of 89% [12]. A comparative study found ResNet50 reached 89.09% accuracy with the HAM10000 dataset, and VGG19 reached 86.21% with the Kaggle dataset, among ResNet50, VGG19, and Vision Transformers with XAI [13]. A comparative study reported that histogram-based local

descriptors with XGBoost achieved 90% accuracy with the Kaggle dataset, using colored LDN [14]. An approach applied transfer learning with VGG16 to diagnose basal cell carcinoma, melanoma, and nevi, training on 13,232 with merging the Kaggle (ISIC) and HAM10000 dataset, and achieved 84.5% testing accuracy, where the authors noted that image quality played a more important role than quantity [15]. Skin cancer detection using a CNN and image processing achieved 85.8% accuracy with the ISIC2018 dataset and InceptionV3 transfer learning, also integrated into a mobile application [16]. Studies have shown the successful implementation of deep learning models in critical tasks such as brain tumor detection using MRI images [17]. The researchers also applied VGG16, a deep learning model, to classify disease detection by image analysis [18].

Most of these studies are commonly trained and evaluated in a publicly accessible dataset in the Kaggle repository, which originated from the ISIC Archive [8]. Across prior studies, deep learning and pre-trained models have consistently improved diagnostic accuracy in skin cancer classification, with reported performance ranging from 84.24% to 90%. Many studies combined deep learning models with transfer learning, fine-tuning, and explainable AI techniques. Despite these advances, recurring challenges include the persistent need for higher classification accuracy, better generalization with reduced overfitting, enhanced model interpretability, and overcoming real-world deployment challenges in clinical environments. This research directly addresses these identified gaps by aiming for robust performance, integrating crucial interpretability, and enabling real-time deployment for improved skin cancer classification.

3. Proposed System: Design and Implementation

This research introduces an interpretable and enhanced transfer learning model for binary skin cancer classification. The system utilizes the strong feature extraction capabilities of the pre-trained VGG16. To adapt the learned representations of VGG16 to medical imagery, only its final convolutional layers are fine-tuned, while the initial layers remain frozen. The high-level abstract extracted features are then fed into a custom-designed classifier head for classification. Augmented Grad-CAM++ is integrated to enhance diagnostic transparency. For practical application, a Flask-based desktop application is developed, offering an interactive interface that supports real-time image analysis, results and its visualization, and report generation.

3.1. Model and System Architecture

An interpretable and enhanced transfer learning model is developed to classify skin cancer as benign or malignant. The architecture of the model is illustrated in Figure 1.

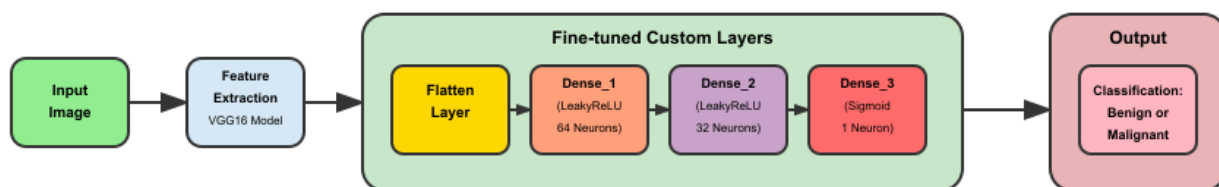


Figure 1. Model Architecture.

The approach employs transfer learning by utilizing the VGG16, a well-established convolutional neural network model, pre-trained on the ImageNet dataset for extracting deep features from input images. The final convolutional block of the VGG16 model is unfrozen and fine-tuned using a labeled dataset of dermoscopic images. This selective fine-tuning enables the model to capture domain-specific features for accurate skin lesion classification, while the frozen layers preserve the general visual features acquired during initial training. Following feature extraction, the classifier head is incorporated with a

flatten layer to reshape the output in a one-dimensional vector. Then it is processed by two dense layers that respectively contain 64 and 32 neurons. The final dense layer outputs binary class probabilities using a sigmoid activation function. The simplicity of the design is what sets the proposed model architecturally apart from other well-known CNN models or transformer-based models. The proposed VGG16-based model is more effective because it relies on a single convolutional block and employs fewer layers, making its construction simpler and more computationally efficient.

Several optimization techniques have been applied to reduce overfitting, ensure stable training, and improve classification accuracy. The LeakyReLU activation function is applied to mitigate the vanishing gradients problem in both dense layers (Dense 1 and Dense 2 in Figure 1). Batch normalization and L2 regularization are applied to ensure stability and reduce overfitting, with a dropout layer further preventing over-reliance on specific neurons. Class weights are dynamically calculated and used during training to reduce the class bias to address the class imbalance problem in the training data (1440 benign samples and 1197 malignant samples) and improve the classification accuracy.

To improve the classification of skin lesions, this study uses a transfer learning approach with the VGG16 model. The architecture of the VGG16 model is illustrated in Figure 2. It consists of five convolutional blocks with 2×2 max-pooling operations after 3×3 filters and ReLU activations to gradually increase feature depth while decreasing spatial dimensions.

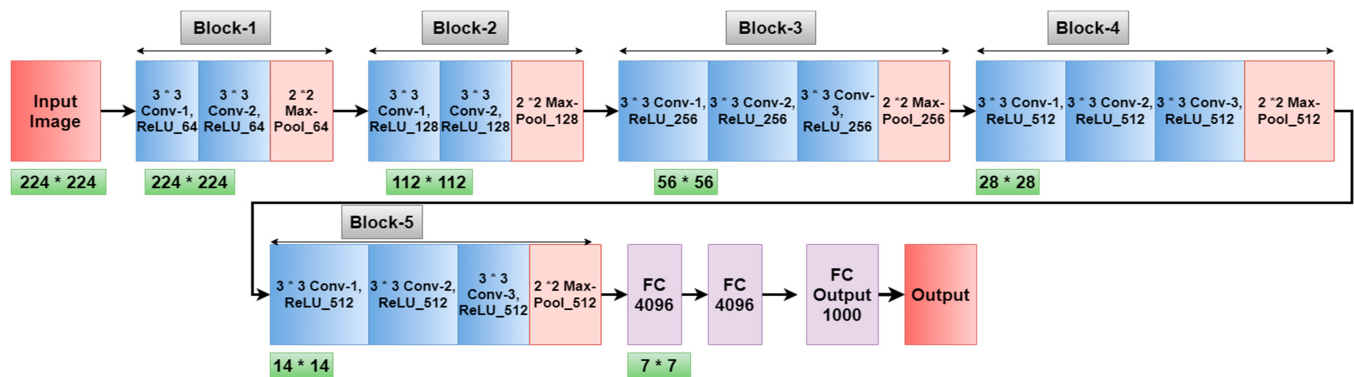


Figure 2. VGG16 Architecture.

The final convolutional block of the VGG16 model is used to extract features in this research. This convolutional block was selected because of its ability to capture abstract high-level features relevant to the classification of skin lesions. By freezing the remaining layers, the model retains the general feature representations learned from the ImageNet dataset while ensuring efficient and stable learning with limited medical data. The extracted features are passed through the custom classifier head tailored for binary classification. This transfer learning strategy provides a balance between the use of deep representational power and the assurance of efficiency, enabling effective classification even with limited domain-specific data.

VGG16 is selected as the feature extractor due to its proven efficiency in image classification and transfer learning capabilities. VGG16 is especially useful for medical imaging tasks due to its very small depth compared to transformer-based models. It is ideal for extracting complex hierarchical patterns, which are essential for the classification of skin cancer. Improved performance reported in a study that surpasses state-of-the-art approaches in skin cancer classification with less labeled data by utilizing transfer learning capabilities of VGG16, which uses pre-trained weights from large datasets such as ImageNet [1]. The feature extraction efficiency of VGG16 allows it to capture complex detailed

patterns in dermoscopic images, making it an excellent choice for skin image analysis. Numerous studies have confirmed its success in medical image analysis, notably in the classification of skin cancer, demonstrating its reliability and strong performance in this area [9,15].

The overall system architecture comprises five main modules comprising image acquisition, web upload, back-end processing, results display, and report generation as visualized in Figure 3.

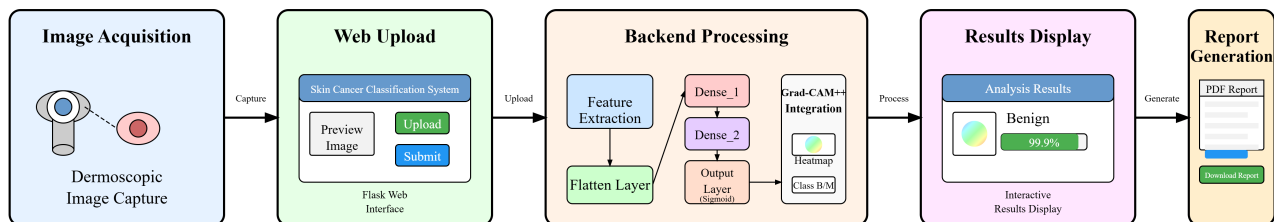


Figure 3. System Architecture.

Within the image acquisition module, dermoscopic images are acquired to maintain visual quality and detail. These images are then uploaded through a Flask-based web interface that facilitates seamless user interaction. The back-end processing module manages all computational tasks, including preprocessing, feature extraction, classification, and heatmap generation. The results display module presents the classification output, confidence scores, an uploaded image preview, and corresponding Grad-CAM++ heatmap in a visually interpretable format. Finally, the Report Generation module allows users to export a detailed PDF report containing diagnostic results, images, confidence levels, input image, and an associated heatmap to support informed clinical decision making.

3.2. Requirement Analysis

The development, deployment, and use of the proposed system require specific hardware and software configurations. The hardware setup included a laptop manufactured by Acer Inc. (New Taipei City, Taiwan), equipped with an Intel Core i5-1135G7 Processor with Intel(R) Iris(R) Xe Graphics, 20 GB of RAM, and sufficient storage to handle data and logs. On the software side, Python 3.10.11 and libraries including TensorFlow 2.19, Keras 2.19, NumPy 2.1.3, OpenCV 4.11.0.86, Matplotlib 3.10.1, Pandas 1.5.3, Seaborn 0.12.2, and Scikit-learn 1.2.2 were used for development within Jupyter Notebook integrated in PyCharm Professional 2024.3.5. In addition, Google Colaboratory was used to train the model with the T4 GPU runtime. The app interface was built with Pywebview (edgehtml) 4.2.1, Flask 3.1.1, HTML5, and CSS3. Pyinstaller 6.14.2 was used to generate the Windows executable file (.exe) of the desktop application.

The requirements of the AI model specify the VGG16 architecture as the base for transfer learning and feature extraction. Additional layers were incorporated for classification along with mechanisms for interpretability. The training was performed applying the Adam optimizer with binary cross-entropy as the loss function, reflecting the binary classification task. Grad-CAM++ integration enables visual explainability, improving the clinical applicability of the model.

For users, hardware essentials include a dermoscope or high-resolution imaging device, computer or smartphone for user interaction, and sufficient storage space to store the user data and downloaded reports. In terms of software, Windows 10 or higher is required to access the interface.

3.3. Implementation

The model was implemented using the Kaggle dataset [8] images and was preprocessed and resized to a uniform dimension of (224, 224, 3), to align with the input specifications of the VGG16 architecture. The dataset was subsequently separated into training, validation, and test sets.

The implementation process began by fine-tuning the final convolutional layers of the VGG16 to capture patterns specific to skin cancer. This selective approach preserved the generalization capabilities of earlier layers and adapted the model to the medical domain. A custom classification head was added after feature extraction, consisting of two dense layers with LeakyReLU activations. Batch normalization, dropout, and L2 regularization were applied to increase convergence, reduce overfitting, and stabilize training.

The Adam optimizer was selected for the proposed model. Adam or Adaptive Moment Estimation is an adaptive learning rate optimization method. The Adam optimizer was selected for its excellent balance of computational efficiency, ease of use, and robust performance in the training of deep learning models. The advantages of the RMSProp and AdaGrad algorithms are combined in Adam. Through the estimation of the first and second moments of the gradients, it determines different adaptive learning rates for each parameter, facilitating faster and more efficient convergence. Adam was selected because of its adaptive learning rate, and the learning rate was adjusted to ensure smooth convergence.

Hyperparameter tuning involves selecting the appropriate number of neurons or units in each dense layer, learning rate, batch size, L2 regularizer threshold, dropout threshold, number of epochs, and loss function. The model was compiled with Adam optimizer with a starting learning rate of 0.001 and the learning rate was adjusted. The optimal learning rate was 0.00005. A range of batch sizes, from 8 to 128, was systematically explored to pinpoint the ideal balance between computational demands and classification accuracy. The trials consistently showed that a batch size of 32 offered this optimal equilibrium. Similarly, when assessing performance across different epoch counts, specifically 8, 10, 16, 20, 30, 50 and the 10 epoch marks consistently yielded peak accuracy without any indication of overfitting. The optimal L2 regularizer threshold was 0.002 for both dense layers (dense layers with 64 and 32 neurons). The optimal dropout threshold for the dense layer with 64 neurons was 0.4, and the dense layer with 32 neurons was 0.3. Binary cross-entropy was utilized as the loss function.

To generate an explanatory heatmap for each prediction, augmented Grad-CAM++ visualization was implemented with binary masking in the back-end of the system. This technique was selected to reduce noise and highlight consistent attention regions. It applies test-time image augmentation, such as flip, rotation, and zoom, especially for generating Grad-CAM++ heatmaps. In the system, each uploaded image is augmented multiple times, and Grad-CAM++ is computed on each augmented image. The generated heatmaps are averaged to produce a more stable and reliable visualization. This interpretability feature, along with the classification results, is presented to users via an interactive desktop app interface. The graphical user interface of the system is shown in Figure 4, and it allows direct engagement, enabling users to upload an image (Figure 4a), get a preview of the uploaded image (Figure 4b), and obtain an instant result with Grad-CAM++ heatmap overlay (Figure 4c), as well as a downloadable report with key details (Figure 4d).

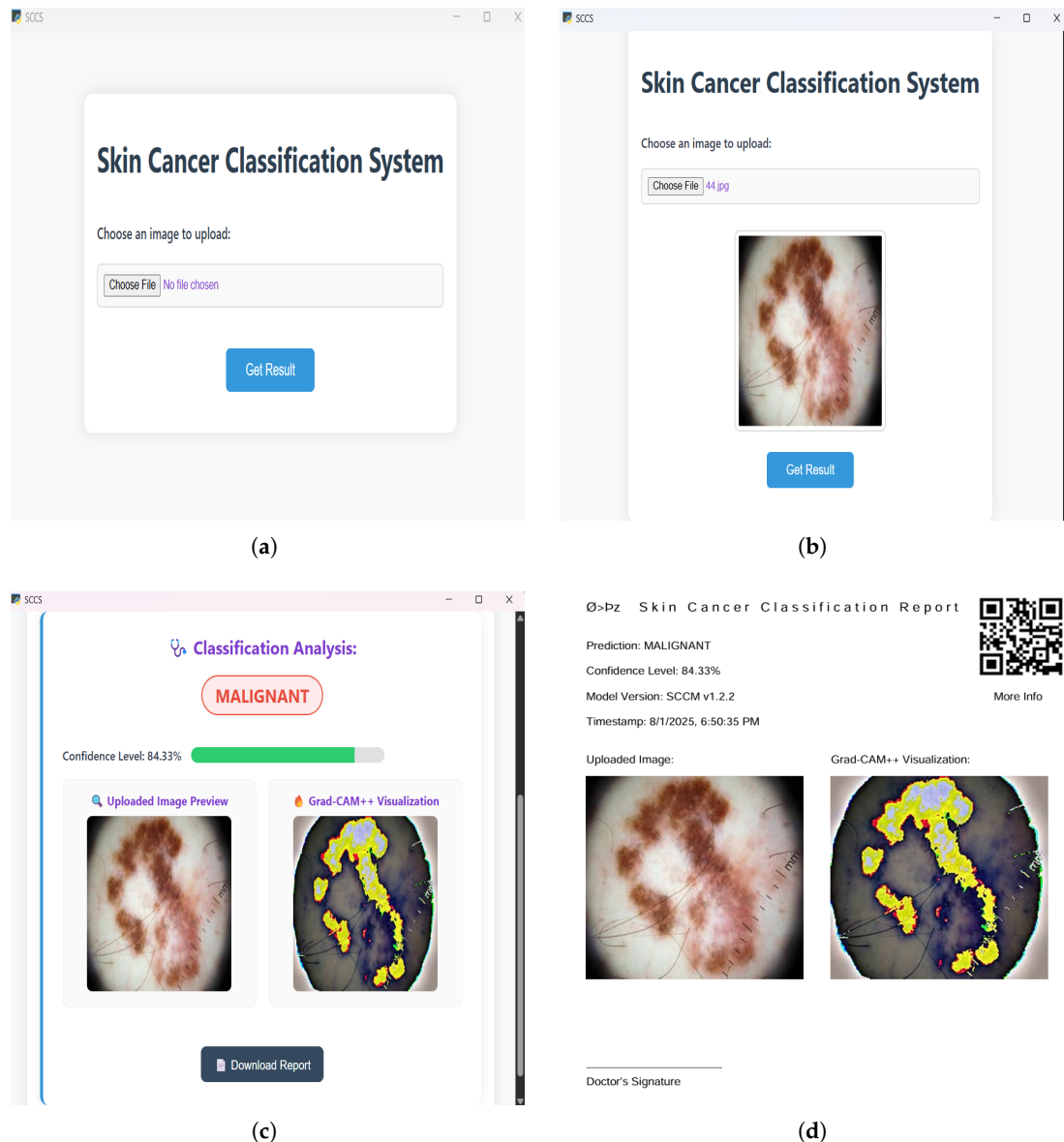


Figure 4. Graphical User Interface of Desktop Application: (a) Upload Image. (b) Preview of the Uploaded Image. (c) Classification Analysis with Heatmap Overlay. (d) Downloaded Report.

4. Result Analysis and Discussion

The proposed model significantly improved the classification results compared to state-of-the-art models and existing deep-learning based approaches in skin cancer classification. An analysis of confusion matrix of the proposed model and a performance metrics comparison between the proposed model and state-of-the-art models was conducted. All models were trained and evaluated on the same dataset [8] to ensure stable and reliable performance evaluations. The hyperparameters for all models were fine-tuned in a consistent manner to ensure a fair and unbiased comparison. In addition, heatmap analysis was conducted to investigate the model's insights into making predictions. Finally, the model's performance was compared in terms of accuracy with existing approaches for binary skin cancer classification, all utilizing the same dataset [8].

4.1. Dataset Specifications

The research utilized a publicly available online dataset in the Kaggle repository named "Skin Cancer– Malignant vs. Benign", originally sourced from the ISIC Archive [8].

The dataset comprises 3297 dermatoscopic images: 1680 malignant and 1617 benign samples. For the model development and evaluation, the dataset was separated into training (73%), validation (7%), and test (20%) sets. Specifically, 2407 samples were allocated for training, and 230 samples for validation, and they were used for hyperparameter tuning and preventing overfitting during model training. The remaining 660 samples were reserved as the independent test set for final model evaluation and comparison. This consistent splitting strategy ensured a robust and unbiased assessment of the performance of the proposed model.

4.2. Training, Validation, and Test

The model accuracy and loss analysis during the training and validation phases confirmed effective learning and generalization, visualized in Figure 5.

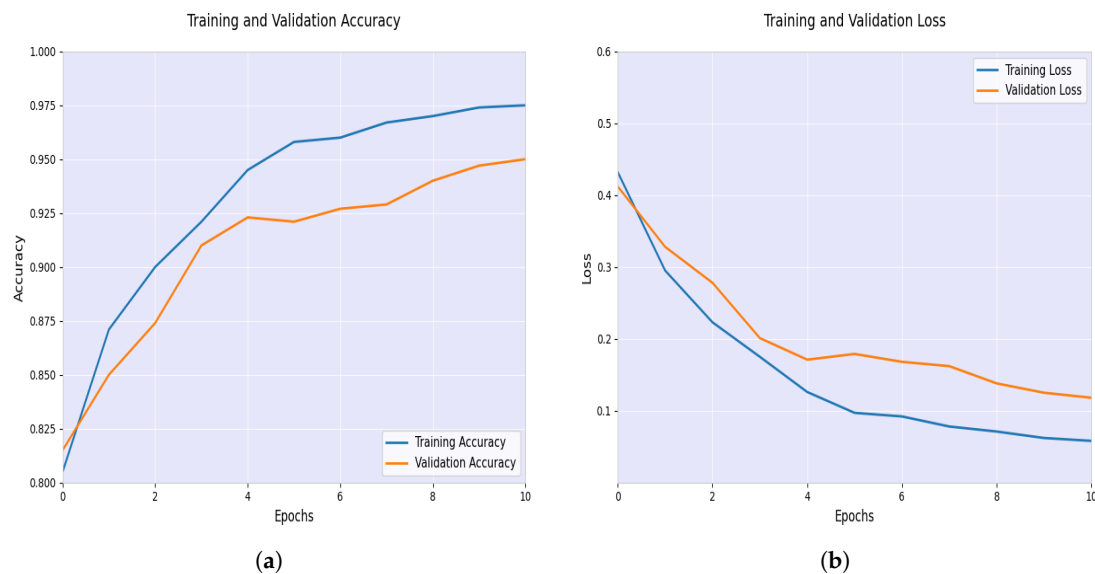


Figure 5. Model Accuracy and Loss during Training and Validation phases: (a) Training and Validation Accuracy. (b) Training and Validation Loss.

As shown in Figure 5a, the training accuracy started approximately 0.805 and uniformly increased with each epoch and reached around 0.975. Also, the validation accuracy started near 0.815 and consistently increased, peaking at about 0.950. The near-identical behavior of the curves reflects effective learning. Training loss started around 0.432 and consistently dropped, hitting roughly 0.058 as represented in Figure 5b. Similarly, validation loss began near 0.412 and steadily decreased, settling at about 0.118. The parallel decrease in both curves indicates stable learning. These results show the strong performance of the model in classifying skin cancer, with the training and validation accuracy achieving high values at the end of the training process. This result confirms the effectiveness of the model in minimizing errors during training while maintaining generalization of the validation data.

During the test phase, the model was evaluated using the independent test set. The test accuracy reached 90.91%, indicating a strong performance in classifying images of skin cancer. A test loss of 0.3570 observed, which reflects low false predictions. Since the accuracy was high while keeping the loss low, this suggests the model has the ability to handle unseen data. This aligns and validates the results observed during training and validation.

Furthermore, to confirm the generalizability of the proposed model in unseen data, it was further evaluated in the independent validation set of a publicly available online

labeled dataset in the Kaggle repository named “Skin Cancer ISIC 2019 & 2020 malignant or benign” [19]. The validation set comprises 1100 dermatoscopic images: 550 malignant and 550 benign samples. In this independent validation set, the validation accuracy was 93.73% with a validation loss of 0.2348. This outcome further demonstrates the effectiveness and robustness of the model.

4.3. Performance Analysis

The confusion matrix of the proposed model is visualized in Figure 6.

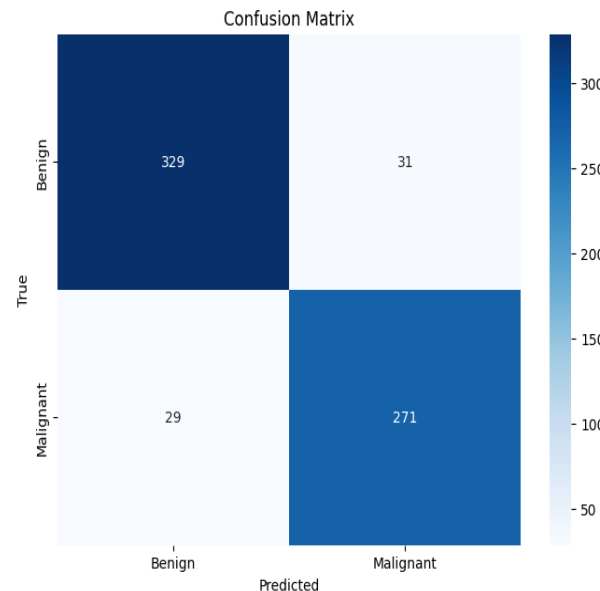


Figure 6. Confusion Matrix of the Proposed Model.

The proposed model correctly classified 329 benign and 271 malignant samples. Only 31 benign and 29 malignant samples were misclassified. Some of the misclassified samples are illustrated in Figure 7, where the benign samples shown in Figure 7a–c were misclassified as malignant, and the malignant samples shown in Figure 7d–f were misclassified as benign.

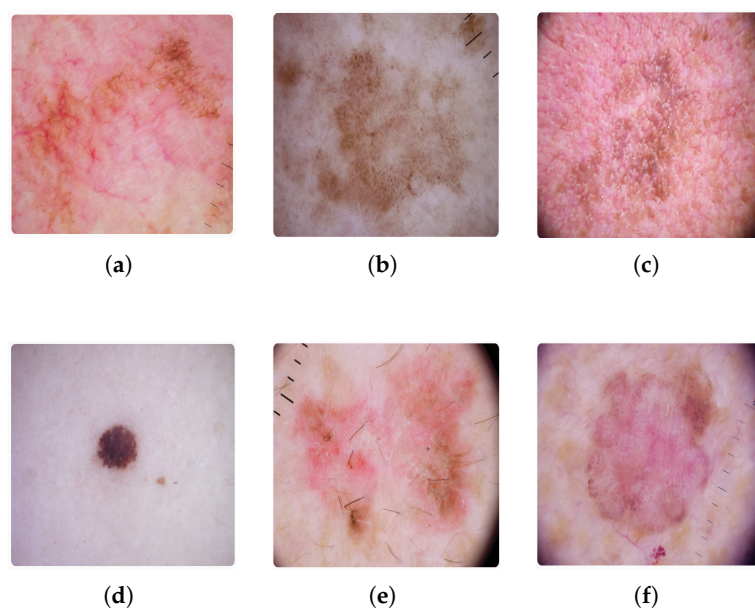


Figure 7. Misclassified Benign and Malignant Images: (a–c) Misclassified Benign Images. (d–f) Misclassified Malignant Images.

The efficacy of the proposed model can be greatly impacted by the misclassification of the samples. The possible reasons behind these misclassification could be insufficient data or data quality issues. Incorrectly classifying benign samples as cancer could result in patients receiving unnecessary therapies. This can lead to intrusive treatments, which could cause mental pain, increased medical costs, and negative impacts. Also, the categorization of malignant samples as benign creates a larger risk since it could result in a delayed diagnosis and treatment, and even decrease their chances of survival. In light of this, future research will concentrate on collecting more data and improving data quality as well as the model to decrease the misclassification rate.

A comparative analysis of performance metrics was conducted between the proposed model and the state-of-the-art models. The hyperparameters for all models were fine-tuned in a consistent manner and used the same dataset [8] to ensure a fair and unbiased comparison. Performance metrics including accuracy, Misclassification Rate, sensitivity (recall), specificity, False Negative Rate (FNR), False Positive Rate (FPR), precision, and F1-score were calculated using Equations (1) to (8) below [1,20].

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}, \quad (1)$$

where TP represents True Positives, TN represents True Negatives, FP represents False Positives, and FN represents False Negatives:

$$\text{Misclassification Rate} = \frac{FP + FN}{TP + TN + FP + FN}, \quad (2)$$

which reflects the proportion of incorrect predictions.

$$\text{Sensitivity (Recall)} = \frac{TP}{TP + FN}, \quad (3)$$

measuring the ability of the model to correctly identify positive cases.

$$\text{Specificity} = \frac{TN}{TN + FP}, \quad (4)$$

which evaluates how well the model identifies negative cases.

$$\text{False Negative Rate (FNR)} = \frac{FN}{TP + FN}, \quad (5)$$

indicating the proportion of positive cases incorrectly predicted as negative.

$$\text{False Positive Rate (FPR)} = \frac{FP}{TN + FP}, \quad (6)$$

denoting the proportion of negative cases incorrectly predicted as positive.

$$\text{Precision (PPV)} = \frac{TP}{TP + FP}, \quad (7)$$

representing the fraction of correctly predicted positive observations among all predicted positives.

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}, \quad (8)$$

which provides a balance between precision and recall.

The comparison between the proposed model and the state-of-the-art models including VGG16, VGG19, ResNet-18, InceptionNet-v4, and AlexNet based on the calculated metrics is summarized in Tables 1–5 below.

Table 1. Comparison of Performance Metrics between the VGG16 Model and the Proposed Model.

Metric	VGG16	Proposed Model	Difference
Accuracy	85.37%	90.91%	+5.54%
Misclassification Rate	14.63%	9.09%	−5.54%
Sensitivity (Recall)	87.42%	90.33%	+2.91%
Specificity	83.61%	91.39%	+7.78%
False Negative Rate	12.58%	9.67%	−2.91%
False Positive Rate	16.39%	8.61%	−7.78%
Precision	82.12%	89.74%	+7.62%
F1-Score	84.67%	90.03%	+5.36%

Table 2. Comparison of Performance Metrics between the VGG19 Model and the Proposed Model.

Metric	VGG19	Proposed Model	Difference
Accuracy	87.61%	90.91%	+3.30%
Misclassification Rate	12.39%	9.09%	−3.30%
Sensitivity (Recall)	80.00%	90.33%	+10.33%
Specificity	94.17%	91.39%	−2.78%
False Negative Rate	20.00%	9.67%	−10.33%
False Positive Rate	5.83%	8.61%	+2.78%
Precision	92.19%	89.74%	−2.45%
F1-Score	85.67%	90.03%	+4.36%

Table 3. Comparison of Performance Metrics between the ResNet-18 Model and the Proposed Model.

Metric	ResNet-18	Proposed Model	Difference
Accuracy	88.51%	90.91%	+2.40%
Misclassification Rate	11.49%	9.09%	−2.40%
Sensitivity (Recall)	88.71%	90.33%	+1.62%
Specificity	88.33%	91.39%	+3.06%
False Negative Rate	11.29%	9.67%	−1.62%
False Positive Rate	11.67%	8.61%	−3.06%
Precision	86.75%	89.74%	+2.99%
F1-Score	87.73%	90.03%	+2.30%

Table 4. Comparison of Performance Metrics between the InceptionNet-v4 Model and the Proposed Model.

Metric	InceptionNet-v4	Proposed Model	Difference
Accuracy	87.31%	90.91%	+3.60%
Misclassification Rate	12.69%	9.09%	−3.60%
Sensitivity (Recall)	86.45%	90.33%	+3.88%
Specificity	88.06%	91.39%	+3.33%
False Negative Rate	13.55%	9.67%	−3.88%
False Positive Rate	11.94%	8.61%	−3.33%
Precision	86.17%	89.74%	+3.57%
F1-Score	86.31%	90.03%	+3.72%

Table 5. Comparison of Performance Metrics between the AlexNet Model and the Proposed Model.

Metric	AlexNet	Proposed Model	Difference
Accuracy	89.10%	90.91%	+1.81%
Misclassification Rate	10.90%	9.09%	−1.81%
Sensitivity (Recall)	91.29%	90.33%	−0.96%
Specificity	87.22%	91.39%	+4.17%
False Negative Rate	8.71%	9.67%	+0.96%
False Positive Rate	12.78%	8.61%	−4.17%
Precision	86.02%	89.74%	+3.72%
F1-Score	88.54%	90.03%	+1.49%

In the majority of key metrics, the proposed model performed better than all other models (VGG16 through AlexNet). The proposed model demonstrates the most improvements over VGG16, with accuracy rising 5.54% points (85.37% to 90.91%) and precision rising

7.62 percentage points. In contrast to VGG19, the proposed model maintains higher specificity and precision while achieving significantly notable sensitivity (+10.33% points) and total accuracy (+3.30% points). All metrics show constant improvements when compared to ResNet-18, with specificity (+3.06% points) and precision (+2.99% points) showing especially significant increases. Comparing InceptionNet-v4 also reveals significant increases, especially in F1-score (+3.72% points) and sensitivity (+3.88% points). With sensitivity (91.29% vs. 90.33%), AlexNet is the closest competitor. The proposed model exceeds it in specificity (+4.17% points), precision (+3.72% points), and overall accuracy (+1.81% points). Overall, the proposed model is the most dependable classifier among all assessed models, achieving the optimal balance of performance measures with continuously high accuracy (90.91%), strong sensitivity (90.33%), good specificity (91.39%), and robust precision (89.74%).

To better understand the model's decision-making process for the classification of skin cancer, a heatmap analysis was performed using Augmented Grad-CAM++, illustrated in Figure 8.

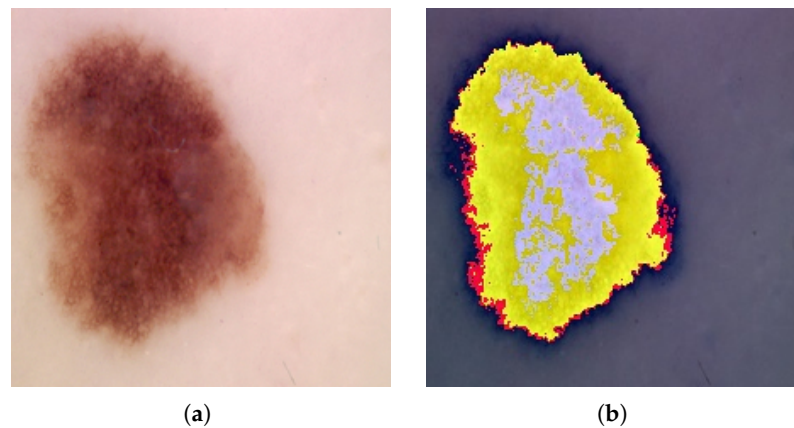


Figure 8. Heatmap Analysis: (a) Input Image. (b) Heatmap Overlay.

The first Figure 8a shows the original dermoscopic input image of skin cancer in the malignant stage, which the proposed model also classified as malignant. The second Figure 8b presents the Augmented Grad-CAM++ heatmap overlay in the input image, which highlights the areas that the model concentrated on while making the predictions. Warmer colors such as white/yellow indicate areas of highest importance, red indicates areas of high importance, green indicates areas of moderate importance, blue indicates areas of low importance, and black indicate areas of no importance. This suggests that the model assigned varying levels of significance to these regions when classifying the lesion as malignant. From the heatmap, it can be observed that the model concentrated on the lesion's irregular borders and pigmentation, which are key clinical indicators of malignancy, aligning well with dermatological understanding, supporting the reliability and interpretability of the model's classification [21]. This enhances the transparency of the AI system, an important factor when deploying models in sensitive applications like preliminary skin cancer diagnosis.

Finally, the model's performance was compared with established studies for binary skin cancer classification in terms of accuracy, all utilizing the same dataset [8], summarized in Table 6.

Table 6. Comparison of Skin Cancer Classification Approaches with the Proposed Model.

Model Architecture	Dataset	Accuracy
Modified VGG16 with data augmentation	Skin cancer: Malignant vs. benign	89.09%
Fine-tuned VGG16 with texture and shape features	Skin cancer: Malignant vs. benign	84.24%
CNN with transfer learning	Skin cancer: Malignant vs. benign	86.65%
Fine-tuned InceptionNet	Skin cancer: Malignant vs. benign	85.94%
Pre-trained deep neural networks (AlexNet, ResNet-18, SqueezeNet, ShuffleNet)	Skin cancer: Malignant vs. benign	89.00% (ResNet-18)
VGG19 with explainable AI	Skin cancer: Malignant vs. benign	86.21%
Histogram-based local descriptors with XGBoost classifier	Skin cancer: Malignant vs. benign	90.00%
VGG16 (last convolutional block) + custom dense classifier + explainable AI (This research)	Skin cancer: Malignant vs. benign	90.91%

The proposed model in this research, which used the last convolutional block of VGG16 for feature extraction combined with a custom dense classifier and explainable AI, achieved an accuracy of 90.91%. This performance exceeded that of several leading approaches in the skin cancer classification domain; for example, an accuracy of 89.09% was achieved by modifying VGG16 architecture and data augmentation [1]. A group of researchers attained 84.24% accuracy through fine-tuning VGG16 and using texture and shape features [9]. Another study showed an accuracy of 86.65% by employing a CNN with transfer learning [10]. A study reported 85.94% accuracy by fine-tuning the InceptionNet model [11]. Pre-trained deep neural networks, such as AlexNet, ResNet-18, SqueezeNet, and ShuffleNet, which were used for skin cancer binary classification, reached the highest accuracy of 89% [12]. A comparative study found that VGG19 reached 86.21% with explainable AI [13]. The proposed approach also outperformed an approach combining histogram-based local descriptors with an XGBoost classifier, which achieved 90% classification accuracy [14]. These comparative results validate the robustness and reliability of the proposed model, demonstrating its potential for accurate and interpretable skin cancer classification.

5. Conclusions

An interpretable and enhanced transfer learning model for the skin cancer binary classification, utilizing the final convolution layer of VGG16 architecture with explainable AI, was developed and evaluated in this research. The model effectively extracted crucial features from dermoscopic images by precisely fine-tuning the final convolutional layers of VGG16 and custom classification layers with advanced activation functions. This architecture showcased substantial performance gains across evaluation metrics, comprising accuracy, precision, recall, specificity, and F1-score. These improvements were observed when the model was compared to the state-of-the-art models and established approaches in skin cancer classification. Interpretability was ensured by implementing Augmented Grad-CAM++, which visualizes the image regions that the model focused on while making predictions, which supports transparency and clinical decision-making. In addition, the model was integrated into an interactive offline desktop application that increased its practical value, providing a real-time skin cancer analysis tool for dermatologists. By acknowledging the system's dependency on dermoscopic images, future works will focus on integrating dermoscope-based IoT devices and further model optimization.

Author Contributions: Conceptualization, M.R.A., F.A.F. and M.G.K.; software/validation, M.R.A. and F.A.F.; formal analysis/investigation, M.R.A., F.A.F.; writing—original draft preparation, M.R.A., F.A.F., J.U. and M.G.K.; writing—review and editing, M.R.A., F.A.F., S.M. and M.G.K.; supervision, J.U., M.G.K. and S.M.; funding acquisition, S.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Multimedia University, Cyberjaya, Selangor, Malaysia (Grant Number: PostDoc(MMUI/240029).

Data Availability Statement: The data supporting the reported results of this study is publicly available and can be accessed in the Kaggle repository at the following link: <https://www.kaggle.com/datasets/fanconic/skin-cancer-malignant-vs-benign> (accessed on 8 June 2025.)

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

Adam: Adam stands for Adaptive Moment Estimation. AI: Artificial Intelligence. CNN: Convolutional Neural Network. DL: Deep Learning. FNR: False Negative Rate. FPR: False Positive Rate. FN: False Negative. FP: False Positive. Grad-CAM: Grad-CAM stands for Gradient-weighted Class Activation Mapping. Grad-CAM++: An enhanced version of the original Grad-CAM technique. PPV: Positive Predictive Value (Precision). SCCM: Skin Cancer Classification Model. SCCS: Skin Cancer Classification System. TN: True Negative. TP: True Positive. VGG16: VGG stands for Visual Geometry Group and 16 refers to the number of layers in the model. XAI: Explainable Artificial Intelligence.

References

1. Anand, V.; Gupta, S.; Altameem, A.; Nayak, S.R.; Poonia, R.C.; Saudagar, A.K.J. An enhanced transfer learning based classification for diagnosis of skin cancer. *Diagnostics* **2022**, *12*, 1628. [CrossRef] [PubMed]
2. Cancer Research UK. Melanoma Skin Cancer Incidence Statistics. Available online: <https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/melanoma-skin-cancer/incidence> (accessed on 18 June 2025).
3. World Health Organization. Ultraviolet Radiation. Available online: <https://www.who.int/news-room/fact-sheets/detail/ultraviolet-radiation> (accessed on 18 June 2025).
4. American Cancer Society. Survival Rates for Melanoma Skin Cancer by Stage. Available online: <https://www.cancer.org/cancer/types/melanoma-skin-cancer/detection-diagnosis-staging/survival-rates-for-melanoma-skin-cancer-by-stage.html> (accessed on 18 June 2025).
5. Ghorbani, M.; Raahemifar, K.; Mahjani, F.; Moradi, F. Early detection of skin cancer using AI: Deciphering dermatology images for melanoma detection. *AIP Adv.* **2024**, *14*, 040701. [CrossRef]
6. Reddy, S.; Shaheed, A.; Patel, R. Artificial intelligence in dermoscopy: Enhancing diagnosis to distinguish benign and malignant skin lesions. *Cureus* **2024**, *16*, e22547. [CrossRef] [PubMed]
7. Wu, Y.; Chen, B.; Zeng, A.; Pan, D.; Wang, R.; Zhao, S. Skin Cancer Classification With Deep Learning: A Systematic Review. *Front. Oncol.* **2022**, *12*, 893972. [CrossRef] [PubMed]
8. Fanconi, C. Skin Cancer: Malignant vs. Benign. Available online: <https://www.kaggle.com/datasets/fanconic/skin-cancer-malignant-vs-benign> (accessed on 18 June 2025).
9. Ibrahim, A.M.; Elbasheir, M.; Badawi, S.; Mohammed, A.; Alalmin, A.F.M. Skin cancer classification using transfer learning by VGG16 architecture (case study on Kaggle dataset). *J. Intell. Learn. Syst. Appl.* **2023**, *15*, 67–75. [CrossRef]
10. Agarwal, K.; Singh, T. Classification of skin cancer images using convolutional neural networks. *arXiv* **2022**, arXiv:2202.00678. [CrossRef]
11. Bazgir, E.; Haque, E.; Maniruzzaman, M.; Hoque, R. Skin cancer classification using Inception Network. *World J. Adv. Res. Rev.* **2024**, *21*, 839–849. ... [CrossRef]
12. Hussein, H.; Magdy, A.; Abdel-Kader, R.F.; Ali, K.A.E. Binary Classification of Skin Cancer using Pretrained Deep Neural Networks. *Suez Canal Eng. Energy Environ. Sci.* **2023**, *1*, 10–14. [CrossRef]
13. Alrabai, A.; Echtioui, A.; Kallel, F. Exploring Pre-Trained Models for Skin Cancer Classification. *Appl. Syst. Innov.* **2025**, *8*, 35. [CrossRef]
14. Yildiz, A. A comparative analysis of skin cancer detection applications using histogram-based local descriptors. *Diagnostics* **2023**, *13*, 3142. [CrossRef] [PubMed]

15. Djaroudib, K.; Lorenz, P.; Bouzida, R.B.; Merzougui, H. Skin cancer diagnosis using VGG16 and transfer learning: Analyzing the effects of data quality over quantity on model efficiency. *Appl. Sci.* **2024**, *14*, 7447. [\[CrossRef\]](#)
16. Othman, S.; Mourad, H. Skin Cancer Detection Using Convolutional Neural Network (CNN). *J. Acs Adv. Comput. Sci.* **2022**, *13*, 42–48. [\[CrossRef\]](#)
17. Nayan, A.A.; Mozumder, A.N.; Haque, M.R.; Sifat, F.H.; Mahmud, K.R.; Azad, A.K.A.; Kibria, M.G. A Deep Learning Approach for Brain Tumor Detection from MRI Images. *Int. J. Electr. Comput. Eng.* **2022**, *13*, 1039–1047. [\[CrossRef\]](#)
18. Akther, J.; Harun-Or-Roshid, M.; Nayan, A.A.; Kibria, M.G. Transfer Learning on VGG16 for the Classification of Potato Leaves Infected by Blight Diseases. In Proceedings of the 2021 Emerging Technology in Computing, Communication and Electronics (ETCCE), Online, 21–23 December 2021. [\[CrossRef\]](#)
19. Ibrahim, S. Skin Cancer ISIC 2019 & 2020 Malignant or Benign. Available online: <https://www.kaggle.com/datasets/sallyibrahim/skin-cancer-isic-2019-2020-malignant-or-benign> (accessed on 10 July 2025).
20. Sathyanarayanan, S.; Tantri, B.R. Confusion matrix-based performance evaluation metrics. *Afr. J. Biomed. Res.* **2024**, *27*, 4023–4031. [\[CrossRef\]](#)
21. Garbe, C.; Amaral, T.; Peris, K.; Hauschild, A.; Arenberger, P.; Basset-Seguín, N.; Bastholt, L.; Bataille, V.; Del Marmol, V.; Dréno, B.; et al. European consensus-based interdisciplinary guideline for melanoma. Part 1: Diagnostics: Update 2022. *Eur. J. Cancer* **2022**, *170*, 236–255. [\[CrossRef\]](#)

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.