Comparative Evaluation of CNN Architectures for Skin Cancer Classification

Taopik Hidayat, Nurul Khasanah, Elly Firasari, Laela Kurniawati, Eni Heni Hermaliani
Faculty of Information Technology, Universitas Nusa Mandiri, Jakarta, Indonesia

***Abstract*—Skin cancer is one of the fastest-growing health problems worldwide. Early and accurate diagnosis is essential for improving treatment success and patient survival. However, many previous studies have focused on single CNN architectures or limited datasets, resulting in models with restricted generalizability. To address this gap, this study presents a comparative evaluation of three deep learning architectures (DenseNet169, MobileNetV2, and VGG19) for automatic classification of benign and malignant skin cancers using dermoscopic digital images. A total of 10,000 images were compiled from three public Kaggle datasets, preprocessed through resizing and data augmentation, and trained using transfer learning based on ImageNet weights. Two data split schemes (60:20:20 and 80:10:10) were applied to assess model robustness. Experimental results show that DenseNet169 achieved the highest test accuracy of 90.7 per cent, while MobileNetV2 was the fastest with an inference time of 16 seconds. These findings highlight the tradeoff between accuracy and computational efficiency and support the use of deep learning models, particularly DenseNet169 and MobileNetV2, in the development of real-time AI-assisted skin cancer diagnostic systems.**

Keywords*—Artificial intelligence; convolutional neural network; deep learning; dermoscopic images; skin cancer classification

I. Introduction

Skin cancer is a common type of cancer that has been increasing rapidly year after year [1]. According to WHO data, each year, there are over two million cases of skin cancer diagnosed, the majority of which are benign, such as actinic keratosis and basal cell carcinoma [2]. Malignant skin cancers, such as melanoma, continue to pose a substantial danger due to their high metastatic potential and death rate [3]. Since the speed at which the lesion can be identified and treated has a significant impact on treatment outcomes, early diagnosis is an essential step in lowering melanoma mortality. Unfortunately, the standard skin cancer detection procedure continues to rely mainly on visual inspection by a dermatologist, which is subjective and requires extensive clinical experience, rendering it prone to errors and inconsistencies [4].

Automation attempts in medical diagnosis have become the subject of significant research, particularly when it comes to medical image processing [5], [6], [7], [8], [9], [10], [11]. The use of deep learning, which excels in extracting visual features and accurately classifying objects, is one of the popular methods [12]. Deep learning can learn visual patterns from picture data from start to finish, eliminating the need for human feature extraction [13]. Deep learning can distinguish tiny distinctions between skin lesion types that are often not visible to the human eye, making it a promising technology to support objective, rapid, and efficient clinical judgments [14], [15], [16].

Several studies in classifying skin lesions in dermatoscopic images have successfully demonstrated using CNN models. Previous studies [17] proposed a hybrid deep learning model to improve the accuracy of skin cancer classification by combining ConvNeXtV2 blocks and a separable self-attention mechanism, especially in differentiating between benign and malignant lesions that share a similar appearance. This model was shown to be more superior than the other ten CNN and ViT models with an accuracy of 93,48% and only 21,92 million parameters, and also efficient for clinical applications. Furthermore, deep learning methods with ResNet, VGG16, and AlexNet models were applied to the HAM10000 dataset for skin cancer classification, with improvements through data augmentation and imbalanced data addition, resulting in accuracies of 92,9%, 98,4%, and 88,3%, respectively, where the improved VGG16 model was selected as the best model to support early detection of skin cancer [18]. Further research [19] developed an automatic skin cancer detection system based on the DenseNet169 model trained using the Skin Cancer: Malignant versus Benign dataset, and successfully achieved a high accuracy of 89.7%, surpassing several conventional models and potentially supporting medical personnel in more accurate and efficient early diagnosis. Six deep learning models, including EfficientNet (B0, B1, B2) and MobileNet (V2, V3-Small, V3-Large), were used in the study [20]. Each model uses the ISIC dataset to classify skin cancer. The best results were obtained from MobileNet-V3-Large with an accuracy of 89.41%, a recall of 90.59%, and an F1-score of 89.53%, accelerating diagnosis and improving accuracy [20]. Another study focused on the use of the MobileNetV2 model for digital image-based skin cancer classification. After training and evaluation, the MobileNetV2 model successfully achieved an accuracy of 85% in distinguishing between cancerous and non-cancerous skin lesion images [21]. In another study [22], several pre-trained models are applied in the deep learning methods, one of which was VGG19. The VGG19 model achieved an accuracy of 87% using the HAM10000 dataset after the augmentation and fine-tuning process, showing quite good performance in classifying skin cancer. While the study [23] enhanced the pre-trained VGG19 model by adding max pooling and dense layers to improve the skin cancer prediction capability. The improved VGG19 model successfully achieves an accuracy of 88% in skin cancer classification.

Referring to the background, the purpose of this study is to evaluate and compare the performance of seven deep learning architectures in classifying skin cancer into two main categories,

namely benign and malignant. The models used are MobileNetV2, InceptionV3, Xception, DenseNet169, ResNet50, VGG16, and VGG19, each of which represents a variation in architectural complexity and computational efficiency. The dataset used was obtained from the Kaggle platform, which consists of 10,000 digital images that have been labeled according to the type of lesion. This study uses a transfer learning approach, where all models are initialized with weights from pre-training on ImageNet, then the medical data is fine-tuned to match its characteristics.

The main purpose of this study is to achieve an optimal deep learning model to classify skin cancer based on evaluation metrics, including recall, precision, accuracy, and F1-score. The results of this study are expected to contribute in the advancement of digital picture development-based clinical decision support systems, especially in the diagnosis of skin cancer. In addition, this study also aims to encourage the use of artificial intelligence in the health sector more widely by considering the aspects of reliability, speed, and efficiency of medical classification systems. It is also expected to provide a thorough grasp of the strengths and the limitations of deep learning architecture in the classification of benign and malignant skin cancer by conducting a comprehensive evaluation of the performance of each model. In addition, the results of this study can be a starting point for the development of an intelligent diagnostic system that is more accurate, easily accessible, and able to support early detection of skin cancer effectively.

The novelty of this study lies in the comprehensive comparison of CNN architectures using multiple public datasets and dual data split schemes to assess both accuracy and inference efficiency in real-world skin cancer classification.

The remainder of this study is organized as follows: Section II reviews related studies on skin cancer classification using deep learning. Section III explains the materials and methods, including dataset composition, preprocessing, and CNN architectures. Section IV presents the experimental results and analysis, while Section V concludes the study with key findings and future directions.

II. Related Works

The previous related studies using digital image processing approaches and deep learning methods for image classification are summarized in Table I. Various deep learning models such as DenseNet169, DenseNet121, MobileNetV2, VGG19, ResNet50, Xception, InceptionV3, and EfficientNetB0 have been applied in these studies. The accuracy results of these models vary, ranging from 85% to 89%.

Table I presents a comparison of the results of several previous studies using a deep learning approach to classify and detect skin cancer. Most studies use a digital image-based approach with popular models including DenseNet, MobileNet, VGG, ResNet, Xception, Inception, and EfficientNet. An accuracy of 89.7% is the highest result using the DenseNet169 model [19], followed by the E-VGG16 model [24] and DenseNet121 [25], which each recorded an accuracy of 89%. Another study [22], [23] showed competitive results using VGG19 and its modified version, with accuracies of 87% and 88%. These results indicate that deep learning models are suitable for skin cancer classification with high accuracy results.

In addition, the MobileNetV2 model was used in two different studies with accuracy results of 85.6% and 85% respectively, indicating the efficiency of this model in a limited computing environment, even with a slight decrease in accuracy [20], [21]. ResNet50 also showed consistency with 87% accuracy in two separate studies [10], [26]. Further research used a combination of features from the Xception and InceptionV3 models to detect Monkeypox skin disease and obtained an accuracy of 85.9%, indicating the potential of the combination architecture approach [27]. Meanwhile, research using EfficientNetB0 recorded an accuracy of 87% [28], showing the superiority of a lightweight yet effective model. Overall, the data in this table indicates that various deep learning models have demonstrated competitive performance and are worth considering for image-based skin disease classification systems. However, few studies have systematically examined how variations in dataset composition and data split schemes affect the performance and generalizability of CNN architectures for skin cancer classification.

TABLE I. Related Works Summary

Author	Task	Method	Acc (%)
R. Pathania et al. [19]	Skin cancer detection	DenseNet169	89.7
O. Sahin et al [20]	Skin cancer classification	MobileNetV2	85.6
D. Moturi et al [21]	Melanoma skin cancer detection	MobileNetV2	85
I. Ahmad et al [22]	Skin cancer detection	VGG19	87
I. Kandhro et al [23]	Skin cancer detection and classification	E-VGG19	88
D. Albashish et al [26]	Melanoma skin cancer classification	ResNet50	87
N. Pratama et al [27]	Monkeypox skin disease detection	Combining feature Xception and InceptionV3	85.9
N. Khasanah et al [10]	Melanoma skin cancer detection	ResNet50	87
S. Matiray et al [28]	Skin cancer detection	EfficientNetB0	87
V. Anand et al [25]	Skin disease classification	DenseNet121	89
S. Mushtaq et al [24]	Skin cancer classification	E-VGG16	89

III. Materials and Methods

In this section, the stages of the methodology applied in the study are described. This study focuses on testing and comparing seven deep learning models with a transfer learning approach, using public data from the Kaggle platform to classify benign and malignant skin cancers. The entire process is designed systematically, starting from the data input stage, pre-processing, model training, to evaluation of results, as shown in the workflow illustration in Fig. 1.

In this study, the primary data used consisted of 10,000 skin cancer images obtained from the Kaggle platform. The initial stages of the study included data pre-processing, which consisted of image resizing and augmentation applications to increase the diversity and quality of training data. Furthermore, this study tested and compared the performance of seven deep learning model architectures, namely, MobileNetV2, InceptionV3, DenseNet169, ResNet50, Xception, VGG16, and VGG19. Fig. 1 provides an overview of the main stages in the research process. In contrast, Fig. 2 shows the complete research stages of this research, which consist of collecting the data, processing the data itself, continuing to the training process with several training models, the evaluation process, and finally drawing conclusions. Fig. 3 provides a detailed visualization of the training and implementation stages of the model in this study, showing an example of the transfer learning application process flow on the seven models.

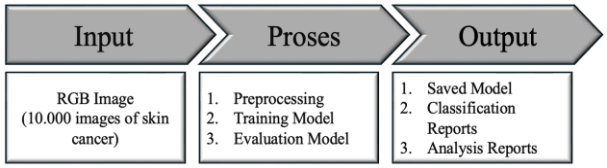

Fig. 1. Overview of the general stages of the research.

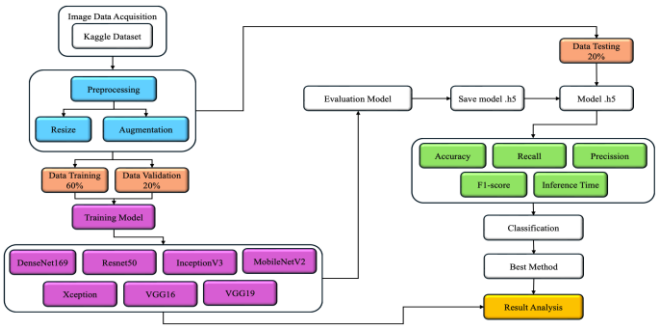

Fig. 2. Flowchart of the research stages.

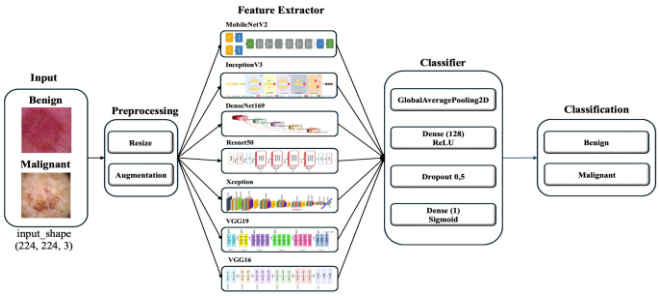

Fig. 3. Proposed model architecture framework.

The initial stage of this research includes collecting skin cancer image data totaling 10,000 images, which are then classified into two main categories, namelyz, benign and malignant. The data is the result of the combination of the three different dataset sources. The first dataset, as many as 3,600 images, was obtained from the Skin Cancer Malignant and Benign Dataset (SCMBD) [29]. Secondly, the source came from the Melanoma Skin Cancer Dataset (MSCD) with 5,000 images [30]. And finally, the third dataset, named the Melanoma Cancer Image Dataset (MCID), consists of 1,400 [31]. All three different datasets are divided into two classes, namely, benign and malignant. Table II shows the details of the data division from each source.

TABLE II. SUMMARY OF THE COMBINED DATASET SOURCE

Class Name	SCMBD Dataset	MSCD Dataset	Additional data MCID	Sub Total
Benign	1,800	2,500	700	5,000
Malignant	1,800	2,500	700	5,000
Total	**3,600**	**5,000**	**1,400**	**10,000**

Training data, validation data, and testing data are the three primary subsets of the skin cancer image data used in this study. To support the experimentation and model evaluation process, two data division schemes are used: the first scheme with a ratio of 60% for training, 20% for validation, and 20% for testing, and an alternative scheme with a ratio of 80%:10%:10%. This division is carried out systematically after all images from the three datasets are combined and standardized in size to 224 × 224 pixels to suit the input needs of the deep learning model [32]. The amount of data in each subset and the division scheme are presented in detail in Table III.

The purpose of using these two division schemes is to ensure that the developed model receives sufficient training data, is optimally validated, and is independently tested so that its performance evaluation is accurate [33]. This approach also allows testing the model's robustness to variations in data proportions and helps find the most effective data sharing configuration in skin cancer image classification. Thus, this data sharing strategy also contributes in reducing evaluation bias and increasing the validity of the overall research results [34].

TABLE III. DATASET DISTRIBUTION

Ratio	Subset	Total data in each subset	Total dataset
Scheme 1 60:20:20	Data training	6,000	10,000
	Data validation	2,000	
	Data testing	2,000	
Scheme 2 80:20:20	Data training	8,000	10,000
	Data validation	1,000	
	Data testing	1,000	

To meet the input dimensions required by all the CNN architectures, the dataset in this study was standardized to 224×224 pixels. To increase the capacity of training data and overcome the limitations of the number of datasets, this study applied an image augmentation technique that was specifically carried out only on a subset of the training data (training set). The augmentation process involves various image transformations, including scaling, rotation, translation, zoom, horizontal flipping, vertical and horizontal position shifts (height and width shift), and brightness adjustment. This augmentation technique not only functions to artificially increase the amount of data but also enriches the variety of visual patterns, which are learned in the training process by the model. With the increasing diversity of augmented data, it is hoped that the model can recognize features more comprehensively and have better generalization capabilities to new data outside the training

dataset [35]. Examples of preprocessing and image augmentation results can be seen in Fig. 4, which shows a visualization of the transformation before the data is used in the training process. Details of the augmentation parameters applied during training are presented in Table IV as a technical reference for implementation, with parameters such as horizontal flip enabled, rotation range of 20 degrees, and zoom, shear, width shift, and height shift range, each valued at 0.2. Meanwhile, Table V shows the amount of augmented data from each dataset division scheme in detail.

TABLE IV. TRAINING DATA AUGMENTATION PARAMETERS

Parameters	Value
Horizontal flip	True
Rotation range	20 degrees
Zoom range	0.2
Shear range	0.2
Width shift range	0.2

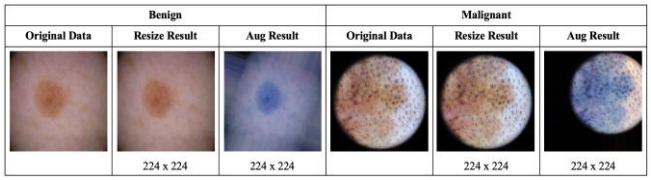

Fig. 4. Example training images from each class.

TABLE V. TRAINING DATASET BEFORE AND AFTER AUGMENTATION

Scheme	Classes	Before augmentation	After augmentation
Scheme 1 60% of dataset	Benign	2,942	5,884
	Malignant	3,058	6,116
Total	**2 classes**	**6,000**	**12,000**
Scheme 2 80% of dataset	Benign	3,998	7,976
	Malignant	4,012	8,024
Total	**2 classes**	**8,000**	**16,000**

Table V shows the amount of data before and after augmentation based on the two dataset division schemes used in this study. In scheme 1, the training data consists of 2,942 benign class images and 3,058 malignant class images. After augmentation, they become 5,884 and 6,116 images, respectively, so that the total training data is 12,000 images. Meanwhile, scheme 2 uses 80% of the entire dataset for training, resulting in 3,998 benign images and 4,012 malignant images, which are multiplied through augmentation to 7,976 and 8,024 images, respectively, with a total of 16,000 training images. The application of this augmentation aims to increase the amount and diversity of data so that the model is better able to recognize visual patterns of both classes of skin cancer accurately.

The training process is carried out using a transfer learning approach using seven CNN architectures, namely MobileNetV2, InceptionV3, Xception, DenseNet169, ResNet50, VGG16, and VGG19. Initial weights that have been previously trained on the ImageNet dataset are initialized on each model, then fine-tuning is carried out so that the model is able to adapt to the visual characteristics of dermatoscopic images of skin cancer [36]. In the training process, the Adam optimization algorithm was used with a learning rate value of 0.001. The performance of each model was monitored periodically based on the accuracy and loss values on the training and validation data in each epoch. The training results were recorded in the form of average training accuracy, average validation accuracy, average training loss, average validation loss, and total training time. After training, the model was saved in an .h5 file and was ready for testing.

The initial stage begins by inputting skin cancer images as input into the model with a size (input shape) of $224 \times 224 \times 3$. The image data consists of two classes, namely benign skin cancer and malignant skin cancer. The training process is carried out on seven CNN model architectures, namely, MobileNetV2, InceptionV3, DenseNet169, ResNet50, Xception, VGG19, and VGG16. The training process is carried out for 20 epochs using the Binary Cross entropy loss function, which is suitable for the binary classification process. A dropout mechanism of 0.5 and an early stopping approach with a 5-epoch patience are used to prevent overfitting. If the model's performance on the validation data does not improve for five consecutive epochs, training will end. A learning rate value of 0.001 as a default of Adam optimizer is also used in the optimization process. At the end of the network (dense layer), 128 neurons are used with the ReLU activation function, and followed by generating the binary output of benign or malignant using the sigmoid activation function in the output layer.

Model evaluation was performed using a subset of test data to measure the final performance of each deep learning architecture in classifying skin cancer images into two categories, namely benign and malignant. Performance evaluation was performed by calculating a number of key classification metrics, such as recall, precision, accuracy, F1-score, and inference time to measure the efficiency of model predictions in the context of real applications [37]. In addition, a confusion matrix was used to summarize the performance of the classification distribution of true (correct) and false (incorrect) predictions from each class, thus offering a more thorough comprehension of the effectiveness of the model in digital image-based classification tasks. The components of the confusion matrix used are True Positive (TP), representing the number of malignant cancer cases that were successfully predicted correctly, while True Negative (TN) shows the number of benign cancer cases that were correctly identified. Meanwhile, False Positive (FP) refers to benign cancer cases that were incorrectly classified as malignant, and False Negative (FN) describes malignant cancer cases that failed to be recognized by the model [37].

The calculation of recall, precision, accuracy, and F1-score follows the formulas outlined in Eq. (1) to Eq. (4). Accuracy provides an overview of the accuracy of the model's predictions. Recall evaluates the model's potential ability to accurately classify cases of malignant cancer, which is very important to avoid the risk of delayed diagnosis. Precision evaluates the accuracy of the model's predictions, avoiding misdiagnosis of healthy patients. Meanwhile, F1-score offers a balanced measure of performance, which particularly relevant when there is an imbalance in the amount of data between classes. It is also known as the harmonic mean of precision and recall. Inference time is also an important indicator in assessing the suitability of a model for application to a real-time detection system in a

clinical context [38]. By applying a combination of these metrics, this study aims to present a thorough evaluation of model performance in digital image-based skin cancer classification.

$$\text{Accuracy}=\frac{TP+TN}{TP+TN+FP+FN}x100 \quad (1)$$

$$\text{Precision}=\frac{TP}{TP+FP}x100 \quad (2)$$

$$\text{Recall}=\frac{TP}{TP+FN}x100 \quad (3)$$

$$\text{F1-score}=2x\frac{\text{Recall}x\text{Precision}}{\text{Recall}+\text{Precision}}x100 \quad (4)$$

IV. Result and Discussion

A. Result

The implementation and comparative evaluation of seven models were carried out using Jupyter Notebook. The Python programming language used in this study is backed by a number of significant libraries, including Keras, NumPy, Pandas, Matplotlib, and TensorFlow, which manage the training and testing stages of the deep learning models. The purpose of this study is to identify the deep learning model that best distinguishes between benign and malignant forms of skin cancer. There are seven types of CNN models tested, namely MobileNetV2, InceptionV3, DenseNet169, ResNet50, Xception, VGG19, and VGG16. All models are trained using an algorithm called Adam, which functions to accelerate and stabilize the training process. Adam was chosen because it is able to adjust its learning rate automatically, so that training is more efficient than other methods, such as RMSprop. A learning rate of 0.001 is often used as an initial value because it is considered stable enough to prevent excessive weight updates but still efficient in reducing loss during training. This value is also very suitable for adaptive optimizers such as Adam, which automatically adjust the learning rate of each parameter. In addition, this learning rate has been proven effective in various studies for models with medium to high complexity and is able to avoid the risk of overshooting the minimum point.

The training process on the seven models was carried out using a fine-tuning strategy, where some early layers of the pre-trained model were kept frozen, while the final layers were retrained using the skin cancer dataset. This strategy allows the model to adjust the general knowledge obtained from large datasets such as ImageNet to the specific characteristics of the dataset used in this study. By applying a dropout rate in the model of 0.5, it is expected to be able to reduce the chance of overfitting while maintaining the ability to capture important details from the training data. Every model has the same initial hyperparameters in order to preserve equality throughout the evaluation process. Table VI shows the results of the training process, including the accuracy, loss, and training time of each model, as a basis for model performance analysis.

Based on the experimental results on seven deep learning architectures with two data splitting scenarios (60:20:20 and 80:10:10), DenseNet169 showed the most outstanding performance. This model managed to achieve the highest validation accuracy, which was 87.07% in the 60:20:20 scenario and increased to 89.57% in the 80:10:10 scenario. In addition, the validation loss value on DenseNet169 was also recorded as the lowest compared to other models, indicating good training stability and generalization ability without symptoms of overfitting. This proves that, despite only the last layer being retrained throughout the fine-tuning procedure, DenseNet169 is able to identify significantly important patterns in the data.

TABLE VI. Comparison of Model Training Result

Model	Split Data	Val Acc (%)	Val Loss (%)
DenseNet169	60:20:20	87.07	29.87
InceptionV3	60:20:20	82.35	39.97
MobileNetV2	60:20:20	83.56	33.63
ResNet50	60:20:20	66.41	64.44
VGG16	60:20:20	81.8	43
VGG19	60:20:20	86.95	31.91
Xception	60:20:20	83.55	38.09
DenseNet169	80:10:10	89.57	23.99
InceptionV3	80:10:10	82.55	36.13
MobileNetV2	80:10:10	85.79	31.29
ResNet50	80:10:10	68.12	61.22
VGG16	80:10:10	84.95	35.82
VGG19	80:10:10	83.89	38.61
Xception	80:10:10	85.69	32.29

Meanwhile, several other models, such as MobileNetV2, Xception, and VGG19, also showed quite good performance with validation accuracy ranging from 83% to 86%, and loss values that are still quite good. These models still have decent potential for use in skin cancer classification, although not as optimal as DenseNet169. On the other hand, ResNet50 consistently recorded the lowest performance, both in terms of validation accuracy and loss rate, in both data scenarios. This indicates that in this study, the ResNet50 architecture is less suitable for the dataset and hyperparameter configuration used. After the training process of the seven deep learning models was completed, each model was saved in .h5 (HDF5) format to preserve the weights and network architecture that had been trained. Saving in this format allows for future reuse of the model without having to repeat the training process from the beginning. The next stage is the testing process for all models using test data that has been previously separated from the training and validation data. The purpose of this testing is to evaluate the final performance of each model on new data that has never been seen before and how well the model can generalize the classification of benign and malignant skin cancers accurately under actual circumstances. The results of testing each model are shown in Table VII.

TABLE VII. COMPARISON OF MODEL TESTING RESULT

Model	Split Data	Test Accuracy (%)	Test Loss (%)	Precision (%)	Recall (%)	F1-score (%)	Inference time
DenseNet169	60:20:20	90.4	9.6	90.4	90.4	90.4	79 s
InceptionV3	60:20:20	86.1	14	86	86.1	86	58 s
MobileNetV2	60:20:20	88	12	88.3	87.7	87.9	16 s
ResNet50	60:20:20	73.7	26.3	74.8	74.2	73.6	55 s
VGG16	60:20:20	85	15	85	85.1	85	194 s
VGG19	60:20:20	88.4	11.6	88.5	88.3	88.4	28 s
Xception	60:20:20	86.7	13.4	86.7	86.5	86.5	78 s
DenseNet169	80:10:10	90.7	9.3	90.7	90.7	90.7	41 s
InceptionV3	80:10:10	83.2	16	84.6	83.2	83.2	25 s
MobileNetV2	80:10:10	87.8	12.2	87.9	87.8	87.8	14 s
ResNet50	80:10:10	72.7	27	72.7	72.7	72.7	40 s
VGG16	80:10:10	82.3	17.7	82.3	82.3	82.3	90 s
VGG19	80:10:10	83.2	16.8	83.2	83.2	83.2	118 s
Xception	80:10:10	86.2	13.8	86.2	86.2	86.2	37 s

Based on the test results, the three best models for skin cancer classification are DenseNet169, MobileNetV2, and VGG19. DenseNet169 shows the best performance with the highest accuracy of 90.7%, as well as equally high recall, precision, and F1-score (90.7%), making it the most reliable and stable model. MobileNetV2 is the most efficient model with the fastest inference time (14 to 16 seconds) and competitive accuracy of up to 88%, suitable for low-resource devices and real-time applications. Meanwhile, VGG19 offers high accuracy (up to 88.4%) and balanced classification metrics, although with a longer inference time (28 to 118 seconds), making it a good choice for systems with high precision requirements and sufficient computing resources.

The analysis results of the three best models based on tests for classifying benign and malignant skin cancer are shown in Fig. 5. A thorough visual comparison of the three models is also provided in Fig. 5 to aid in the study and make it easier to recognize and comprehend the variations in model performance. Based on the results presented in Fig. 5, DenseNet169 (80:10:10) shows the best overall performance with an accuracy value of 90.7%, accompanied by a low loss value (9.3%) and balanced recall, precision, and F1-score scores. The next two best models are VGG19 (60:20:20) and MobileNetV2 (60:20:20), with accuracies of 88.4% and 88.0%, respectively. Considering pure accuracy, the three best models in order are DenseNet169 (80:10:10), VGG19 (60:20:20), and MobileNetV2 (60:20:20). However, if time efficiency is an important factor, MobileNetV2 is worth choosing as a superior alternative model.

B. Discussion

A variety of evaluation metrics of the five models included in this study were analyzed in order to assess the pre-trained CNN models. Several important steps were implemented, including data augmentation to increase the number of samples, and the application of class weighting in the loss function during training to give greater weight to underrepresented classes. This strategy was designed to ensure that the model is able to learn the data without bias towards the majority class. In addition, the balanced batch sampling technique was used to keep the class distribution proportional in each training batch, which helps reduce the impact of class imbalance. The validation and evaluation process was carried out thoroughly using F1-score, recall, precision, and its confusion matrix metrics to ensure prediction accuracy and provide a comprehensive understanding of the strengths and limitations of each CNN model in classifying benign and malignant skin cancers.

Performance optimization was performed on three superior models through the application of fine-tuning techniques. Retraining a number of the model's final layers with a low learning rate and regularization technique reduces the possibility of overfitting. Fine-tuning was applied to DenseNet169, MobileNetV2, and VGG16 by opening some of the final layers, while the initial layers were kept frozen so that the pre-training knowledge was maintained. During the retraining process, the Adam optimizer was used with a standard learning rate value of 0.001 to ensure that the weight update process was stable and efficient.

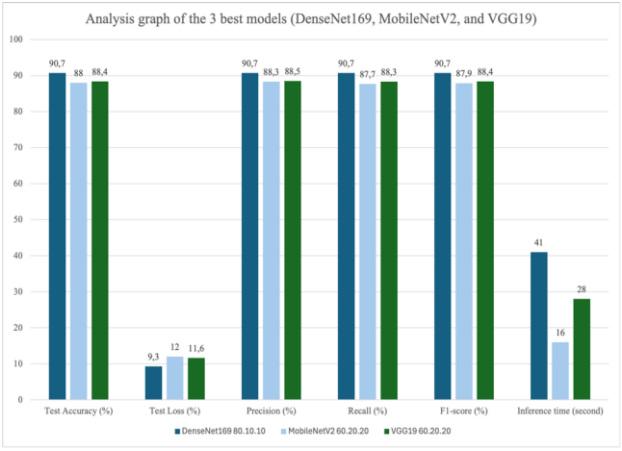

Fig. 5. Analysis graph of the three best models (DenseNet169, MobileNetV2, and VGG19).

This section discusses the best performance obtained from the research results conducted, namely the DenseNet169, MobileNetV2, and VGG19 models. The first result from DenseNet169 produced a test accuracy of 90.7% and a loss of 9.3%. Fig. 6 shows the results with testing data, while Fig. 7 shows the confusion matrix of the DenseNet169 model.

Classification Report:	precision	recall	f1-score	support
benign	0.917	0.893	0.905	495
malignant	0.898	0.921	0.909	505
accuracy			0.907	1000
macro avg	0.907	0.907	0.907	1000
weighted avg	0.907	0.907	0.907	1000

Fig. 6. Classification report of DenseNet169 model.

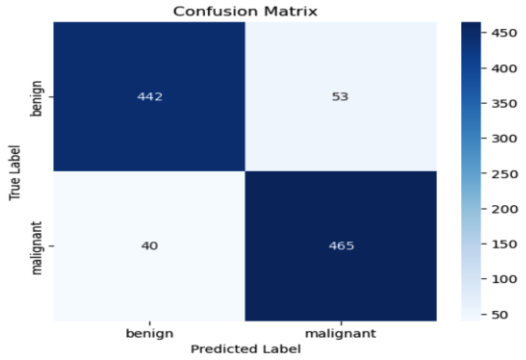

Fig. 7. Confusion matrix of DenseNet169 model.

Based on Fig. 6, the results of the DenseNet169 model show a very good performance in classifying two classes, namely benign and malignant. The precision value, recall, and F1-score of the benign class are 91.7%, 89.3%, and 90.5%, respectively, whereas the malignant class shows 89.8%, 92.1%, and 90.9%. The overall accuracy of the model reaches 90.7%, with macro and weighted average values (macro avg and weighted avg) also consistent at 90.7%, indicating that the model can maintain a balance of performance between classes without significant bias. This reflects how well the model was able to generalize to the test data.

Second, the VGG19 model showed quite good classification performance with an overall accuracy of 88.4%. In the benign class, the model produced a precision value of 87.0% and a recall of 91.2%, with an F1-score reaching 89.1%, which indicates how well the model's ability can identify most benign cases. Meanwhile, in the malignant class, the precision was recorded as higher at 90.1%, but the recall was slightly lower at 85.5%, resulting in an F1-score of 87.7%. This performance reflects that the model tends to be more careful in detecting malignant cases, but is quite effective in reducing false positive predictions. The macro and weighted average values of recall, precision, and F1-score are in the range of 88.3% to 88.6%, respectively, reflecting the stability and balance of classification between classes. Overall, the VGG19 model has good potential for use in image-based skin cancer detection systems with quite competitive performance.

As seen in Fig. 8, it shows the classification report for the VGG19 model, while Fig. 9 shows its confusion matrix. Fig. 8 presents evaluation metrics such as recall, precision, and f1-score for each class (benign and malignant), while Fig. 9 depicts the distribution of correct and incorrect predictions in the form of a confusion matrix. Both figures provide a comprehensive visual representation of the performance of the VGG19 model in classifying skin cancer images.

Classification Report:	precision	recall	f1-score	support
benign	0.870	0.912	0.891	1032
malignant	0.901	0.855	0.877	966
accuracy			0.884	1998
macro avg	0.886	0.883	0.884	1998
weighted avg	0.885	0.884	0.884	1998

Fig. 8. Classification report of the VGG19 model.

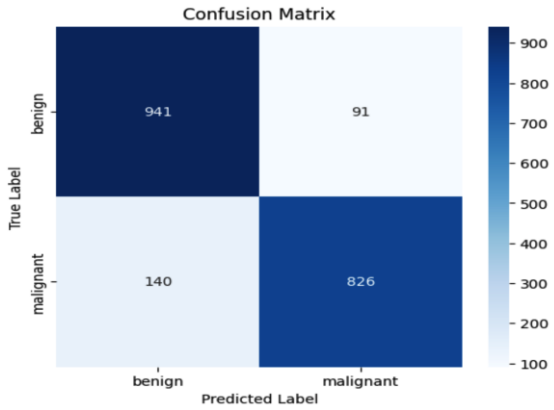

Fig. 9. Confusion matrix of the VGG19 model.

And at last, MobileNetV2, produces a test accuracy of 88% and a loss of 12%. The results of the MobileNetV2 model with testing data and its confusion matrix can be seen in Fig. 10 and Fig. 11, respectively. The MobileNetV2 model shows good performance in classifying benign and malignant categories, with an overall accuracy of 88.0%. The benign class has a precision value of 85.7%, a recall of 92.6%, and an F1-score of 89.0%, indicating that it is quite efficient in identifying most benign cases, although there are several false predictions. In contrast, in the malignant class, the precision value is recorded as higher, namely 91.0%, but with a lower recall of 82.9%, indicating that although the positive predictions for this class are quite accurate, the model still misses a number of malignant cases. The macro and weighted average values of recall, precision, and F1-score range from 87.8% to 88.4%, reflecting a relatively balanced model performance between classes. Despite some disparity in detection between classes, the model still shows good potential for use in general skin cancer image classification.

Based on the evaluation results of the three deep learning models tested, it can be concluded that DenseNet169 is the model with the best performance in skin cancer image classification. With a test accuracy of 90.7% and a loss value of only 9.3%, this model shows a very good balance between recall, precision, and F1-score for both classes, namely benign and malignant. The high macro and weighted average values of 90.7% indicate that this model is able to fairly classify both classes without significant bias. These results indicate that DenseNet169 is not only accurate but also stable and reliable in handling data variations, making it the main choice for a digital image-based skin cancer classification system.

```
Classification Report:
              precision    recall  f1-score   support

      benign       0.857     0.926     0.890      1051
   malignant       0.910     0.829     0.868       949

    accuracy                         0.880      2000
   macro avg       0.884     0.878     0.879      2000
weighted avg       0.882     0.880     0.880      2000
```

Fig. 10. Classification report of MobileNetV2 model.

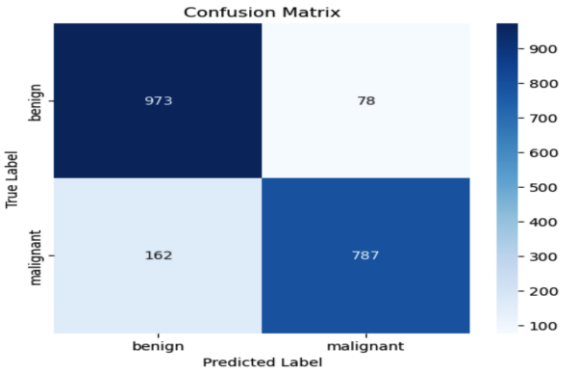

Fig. 11. Confusion matrix of MobileNetV2 model.

The VGG19 and MobileNetV2 models also demonstrate competitive performance, achieving accuracies of 88.4% and 88.0%, respectively. VGG19 exhibits higher recall for the benign class, while MobileNetV2 attains superior precision for the malignant class. Although a minor recall imbalance exists between the two classes, both models deliver stable and consistent performance. MobileNetV2, in particular, shows an advantage in inference efficiency, making it suitable for deployment on resource-constrained devices. Overall, all three models exhibit strong potential for integration into computer-aided skin cancer diagnostic systems. DenseNet169 remains the most robust and accurate architecture, while MobileNetV2 offers a lightweight yet effective alternative for real-time or embedded applications. While this study primarily focuses on classification accuracy, future research may explore explainable AI (XAI) methods such as Grad-CAM to visualize feature attention on dermoscopic images, thereby improving model interpretability and clinical trust. To further validate the proposed approach, Table VIII compares the results of this study with those reported in previous works.

TABLE VIII. COMPARISON OF MODEL EVALUATION RESULTS WITH PREVIOUS STUDIES

Proposed method	Acc (%) This Study	Acc (%) [19]	Acc (%) [20]	Acc (%) [21]	Acc (%) [22]	Acc (%) [23]
Densenet169	90.7	89.7	-	-	-	-
VGG19	88.4	-	-	-	87	88
MobileNetv2	88	-	85.6	85	-	-

V. CONCLUSION

This study evaluated three deep learning architectures, namely DenseNet169, MobileNetV2, and VGG19, for automatic skin cancer classification using digital dermatoscopic images. The dataset combined three public Kaggle sources containing 10,000 benign and malignant cases, evaluated under two data-split schemes (60:20:20 and 80:10:10). Among the models, DenseNet169 achieved the highest accuracy of 90.7%, demonstrating superior robustness and balanced precision-recall performance. MobileNetV2 and VGG19 also showed competitive accuracy (88.0 to 88.4%) with consistent reliability, where MobileNetV2 excelled in inference efficiency, making it suitable for mobile or real-time clinical applications.

These findings confirm the potential of deep learning-based approaches to enhance early skin cancer detection. Each model offers distinct advantages that can be adapted to various healthcare environments depending on system requirements and resource availability. However, this study is limited by dataset diversity in terms of ethnicity, lighting, and geographic distribution, and it does not yet incorporate advanced ensemble learning, lesion segmentation, or explainable AI techniques. Future work will focus on improving generalization through multi-source data integration, exploring model interpretability via XAI methods such as Grad-CAM, and developing lightweight web- or mobile-based diagnostic tools for real-world deployment.

Overall, the results demonstrate that CNN-based architectures, particularly DenseNet169, can significantly contribute to advancing medical image analysis and AI-assisted dermatological diagnostics. In clinical practice, such models can accelerate the diagnostic process, reduce misclassification risks, and improve early detection outcomes. Rather than replacing dermatologists, these systems are designed to support clinical decision-making, helping to improve patient recovery rates and the overall quality of healthcare services.

ACKNOWLEDGMENT

This research was funded by a research grant from the Ministry of Education, Science, and Technology of the Republic of Indonesia. The authors would like to express their sincere gratitude to the Ministry for the financial support that made this study possible. The authors also thank N. Khasanah for her valuable assistance during the research implementation, and L. Kurniawati for her contributions to preparing and organizing the manuscript. Special appreciation is also extended to E. Firasari and E. H. Hermaliani for their help in finalizing the study for publication.

REFERENCES

[1] A. H. Roky et al., "Overview of skin cancer types and prevalence rates across continents," Cancer Pathogenesis and Therapy, vol. 3, no. 2, pp. 89–100, Mar. 2025, doi: 10.1016/j.cpt.2024.08.002.

[2] A. Kurva, M. Korikani, V. Mohan, and R. K. Kancha, "Skin Cancer," in Biomedical Aspects of Solid Cancers, Singapore: Springer Nature Singapore, 2024, pp. 235–252. doi: 10.1007/978-981-97-1802-3_21.

[3] Z. G. Attal et al., "Advanced and Metastatic Non-Melanoma Skin Cancer. Epidemiology, Risk Factors, Clinical Features, and Treatment Options," Biomedicines, vol. 12, no. 7, p. 1448, Jun. 2024, doi: 10.3390/biomedicines12071448.

[4] S. Arige, L. R. Atmakuri, R. Shaik, M. Gude, V. K. Ghanta, and R. Alluri, "Digital Dermoscopy: Advancements in Skin Cancer Diagnosis and Monitoring," Biomedical and Pharmacology Journal, vol. 18, no. December Spl Edition, pp. 33–43, Jan. 2025, doi: 10.13005/bpj/3071.

[5] A. El Mrabet, M. Benaly, I. Alihamidi, B. Kouach, L. Hlou, and R. El Gouri, "Enhancing Early Detection of Skin Cancer in Clinical Practice with Hybrid Deep Learning Models," Engineering, Technology & Applied Science Research, vol. 15, no. 2, pp. 20927–20933, Apr. 2025, doi: 10.48084/etasr.9753.

[6] S. Haque, F. Ahmad, V. Singh, D. M. Mathkor, and A. Babegi, "Skin Cancer Detection Using Deep Learning Approaches," Cancer Biother Radiopharm, vol. 40, no. 5, pp. 301–312, Jun. 2025, doi: 10.1089/cbr.2024.0161.

[7] N. Merlina, A. Prasetio, I. Zuniarti, N. A. Mayangky, D. N. Sulistyowati, and F. Aziz, "Deep CNN Models for Detecting Cervical Cancer in Pap Smear Images," TEM Journal, pp. 1073–1083, May 2025, doi: 10.18421/TEM142-09.

[8] N. Merlina, A. Prasetio, I. Zuniarti, N. Almira Mayangky, D. Nur Sulistyowati, and F. Aziz, "Improving Early Detection of Cervical Cancer Through Deep Learning-Based Pap Smear Image Classification," Journal of Applied Data Sciences, vol. 6, no. 2, pp. 952–968, May 2025, doi: 10.47738/jads.v6i2.576.

[9] E. Firasari, N. Khasanah, F. L. D. Cahyanti, D. N. Kholifah, U. Khultsum, and F. Sarasati, "Performance Evaluation of ResNet50 and MobileNetV2 in Skin Cancer Image Classification with Various Optimizers," in 2024 International Conference on Information Technology Research and Innovation (ICITRI), IEEE, Sep. 2024, pp. 376–380. doi: 10.1109/ICITRI62858.2024.10698943.

[10] N. Khasanah and M. N. Winnarto, "Application of Deep Learning with ResNet50 for Early Detection of Melanoma Skin Cancer," Journal Medical Informatics Technology, pp. 16–20, Mar. 2024, doi: 10.37034/medinftech.v2i1.31.

[11] Daniati Uki Eka Saputri, Nurul Khasanah, F. Aziz, and Taopik Hidayat, "Enhancing Skin Cancer Classification Using Optimized InceptionV3 Model," Journal Medical Informatics Technology, pp. 65–69, Sep. 2023, doi: 10.37034/medinftech.v1i3.14.

[12] X. Zhao, L. Wang, Y. Zhang, X. Han, M. Deveci, and M. Parmar, "A review of convolutional neural networks in computer vision," Artif Intell Rev, vol. 57, no. 4, p. 99, Mar. 2024, doi: 10.1007/s10462-024-10721-6.

[13] J. A. AYENI, "Convolutional Neural Network (CNN): The architecture and applications," Applied Journal of Physical Science, vol. 4, no. 4, pp. 42–50, Dec. 2022, doi: 10.31248/AJPS2022.085.

[14] M. M. Musthafa, M. T R, V. K. V, and S. Guluwadi, "Enhanced skin cancer diagnosis using optimized CNN architecture and checkpoints for automated dermatological lesion classification," BMC Med Imaging, vol. 24, no. 1, p. 201, Aug. 2024, doi: 10.1186/s12880-024-01356-8.

[15] C. Kavitha, S. Priyanka, M. P. Kumar, and V. Kusuma, "Skin Cancer Detection and Classification using Deep Learning Techniques," Procedia Comput Sci, vol. 235, pp. 2793–2802, 2024, doi: 10.1016/j.procs.2024.04.264.

[16] K. Nawaz et al., "Skin cancer detection using dermoscopic images with convolutional neural network," Sci Rep, vol. 15, no. 1, p. 7252, Mar. 2025, doi: 10.1038/s41598-025-91446-6.

[17] B. Ozdemir and I. Pacal, "A robust deep learning framework for multiclass skin cancer classification," Sci Rep, vol. 15, no. 1, p. 4938, Feb. 2025, doi: 10.1038/s41598-025-89230-7.

[18] D. S. Hieu, D. T. Phuc, L. N. Ton, T. C. Hung, and N. C. Cuong, "Improving Architectures of VGG16, AlexNet, and ResNet50 Models for Skin Cancer Classification," 2025, pp. 562–571. doi: 10.1007/978-3-031-90194-2_39.

[19] R. Pathania and P. Behki, "Skin Cancer Detection Using Deep Learning," in 2024 Sixth International Conference on Computational Intelligence and Communication Technologies (CCICT), IEEE, Apr. 2024, pp. 568–575. doi: 10.1109/CCICT62777.2024.00095.

[20] O. Şahin and M. S. Yıldırım, "Performance Comparison of Deep Learning Architectures for Skin Cancer Classification," in 2024 8th International Artificial Intelligence and Data Processing Symposium (IDAP), IEEE, Sep. 2024, pp. 1–8. doi: 10.1109/IDAP64064.2024.10710839.

[21] D. Moturi, R. K. Surapaneni, and V. S. G. Avanigadda, "Developing an efficient method for melanoma detection using CNN techniques," J Egypt Natl Canc Inst, vol. 36, no. 1, p. 6, Feb. 2024, doi: 10.1186/s43046-024-00210-w.

[22] I. Ahmad, B. S. Alsulami, and F. Alqurashi, "Enhancing Skin Cancer Detection with Transfer Learning and Vision Transformers," International Journal of Advanced Computer Science and Applications, vol. 15, no. 10, 2024, doi: 10.14569/IJACSA.2024.01510104.

[23] I. A. Kandhro et al., "Performance evaluation of E-VGG19 model: Enhancing real-time skin cancer detection and classification," Heliyon, vol. 10, no. 10, p. e31488, May 2024, doi: 10.1016/j.heliyon.2024.e31488.

[24] S. Mushtaq and O. Singh, "A deep learning based architecture for multi-class skin cancer classification," Multimed Tools Appl, vol. 83, no. 39, pp. 87105–87127, Jul. 2024, doi: 10.1007/s11042-024-19817-1.

[25] V. Anand, S. Gupta, S. R. Nayak, D. Koundal, D. Prakash, and K. D. Verma, "An automated deep learning models for classification of skin disease using Dermoscopy images: a comprehensive study," Multimed Tools Appl, vol. 81, no. 26, pp. 37379–37401, Nov. 2022, doi: 10.1007/s11042-021-11628-y.

[26] D. Albashish, N. Almansour, A. Abdullah, H. M. J. Mustafa, M. R. AlSayyed, and O. Alrashdan, "Design an Ensemble Pretrained Deep Learning Model for Classification of Melanoma Skin Cancer Images," in 2025 1st International Conference on Computational Intelligence Approaches and Applications (ICCIAA), 2025, pp. 1–7.

[27] N. R. Pratama, D. R. I. M. Setiadi, I. Harkespan, and A. A. Ojugo, "Feature Fusion with Albumentation for Enhancing Monkeypox Detection Using Deep Learning Models," Journal of Computing Theories and Applications, vol. 2, no. 3, pp. 427–440, Feb. 2025, doi: 10.62411/jcta.12255.

[28] S. Matiray and L. K. Singh, "Deep Learning for Early Skin Cancer Detection: A Comparative Study on Hybrid CNN Models," 2025. doi: 10.3233/ATDE250013.

[29] C. Fanconi, "Skin Cancer: Malignant vs. Benign," Kaggle.com. Accessed: Jun. 11, 2025. [Online]. Available: https://www.kaggle.com/datasets/fanconic/skin-cancer-malignant-vs-benign.

[30] M. Hasnain Javid, "Melanoma Skin Cancer Dataset of 10000 Images," Kaggle.com. Accessed: Jun. 11, 2025. [Online]. Available: https://www.kaggle.com/datasets/hasnainjaved/melanoma-skin-cancer-dataset-of-10000-images.

[31] B. Mittal, "Melanoma Cancer Image Dataset," Kaggle.com. Accessed: Jun. 11, 2025. [Online]. Available: https://www.kaggle.com/datasets/bhaveshmittal/melanoma-cancer-dataset.

[32] A. Mahbod, N. Saeidi, S. Hatamikia, and R. Woitek, "Evaluating pre-trained convolutional neural networks and foundation models as feature extractors for content-based medical image retrieval," Eng Appl Artif Intell, vol. 150, p. 110571, Jun. 2025, doi: 10.1016/j.engappai.2025.110571.

[33] H. Bichri, A. Chergui, and M. Hain, "Investigating the Impact of Train / Test Split Ratio on the Performance of Pre-Trained Models with Custom Datasets," International Journal of Advanced Computer Science and Applications, vol. 15, no. 2, 2024, doi: 10.14569/IJACSA.2024.0150235.

[34] P. Georgiadis, E. V. Gkouvrikos, E. Vrochidou, T. Kalampokas, and G. A. Papakostas, "Building Better Deep Learning Models Through Dataset Fusion: A Case Study in Skin Cancer Classification with Hyperdatasets," Diagnostics, vol. 15, no. 3, p. 352, Feb. 2025, doi: 10.3390/diagnostics15030352.

[35] Y. Zhong, W. Zhou, and Z. Wang, "A Survey of Data Augmentation in Domain Generalization," Neural Process Lett, vol. 57, no. 2, p. 34, Mar. 2025, doi: 10.1007/s11063-025-11747-9.

[36] A. Yilmaz, G. Gencoglan, R. Varol, A. A. Demircali, M. Keshavarz, and H. Uvet, "MobileSkin: Classification of Skin Lesion Images Acquired Using Mobile Phone-Attached Hand-Held Dermoscopes," J Clin Med, vol. 11, no. 17, p. 5102, Aug. 2022, doi: 10.3390/jcm11175102.

[37] S. Sathyanarayanan, "Confusion Matrix-Based Performance Evaluation Metrics," African Journal of Biomedical Research, pp. 4023–4031, Nov. 2024, doi: 10.53555/AJBR.v27i4S.4345.

[38] H. E. C. da Silva et al., "The use of artificial intelligence tools in cancer detection compared to the traditional diagnostic imaging methods: An overview of the systematic reviews," PLoS One, vol. 18, no. 10, p. e0292063, Oct. 2023, doi: 10.1371/journal.pone.0292063.