



Article

Multi-Class Skin Cancer Classification Using Vision Transformer Networks and Convolutional Neural Network-Based Pre-Trained Models

Muhammad Asad Arshed ^{1,2,*}, Shahzad Mumtaz ³ , Muhammad Ibrahim ², Saeed Ahmed ¹ ,
Muhammad Tahir ^{4,5} and Muhammad Shafi ⁶

¹ School of Systems and Technology, University of Management and Technology, Lahore 54770, Pakistan; saeed.ahmed@umt.edu.pk

² Department of Computer Science, The Islamia University of Bahawalpur, Bahawalpur 63100, Pakistan; muhammad.ibrahim@iub.edu.pk

³ Department of Data Science, The Islamia University of Bahawalpur, Bahawalpur 63100, Pakistan; shahzadmumtaz22@gmail.com

⁴ Department of Electrical and Computer Engineering, University of Manitoba, Winnipeg, MB R3T 5V6, Canada; m.tahir@umanitoba.ca

⁵ Department of Computer Science, Abdul Wali Khan University, Mardan 23200, Pakistan

⁶ Faculty of Computing and Information Technology, Sohar University, Sohar 311, Oman; mshafi@su.edu.om

* Correspondence: muhammadasadarshed@gmail.com

Abstract: Skin cancer, particularly melanoma, has been recognized as one of the most lethal forms of cancer. Detecting and diagnosing skin lesions accurately can be challenging due to the striking similarities between the various types of skin lesions, such as melanoma and nevi, especially when examining the color images of the skin. However, early diagnosis plays a crucial role in saving lives and reducing the burden on medical resources. Consequently, the development of a robust autonomous system for skin cancer classification becomes imperative. Convolutional neural networks (CNNs) have been widely employed over the past decade to automate cancer diagnosis. Nonetheless, the emergence of the Vision Transformer (ViT) has recently gained a considerable level of popularity in the field and has emerged as a competitive alternative to CNNs. In light of this, the present study proposed an alternative method based on the off-the-shelf ViT for identifying various skin cancer diseases. To evaluate its performance, the proposed method was compared with 11 CNN-based transfer learning methods that have been known to outperform other deep learning techniques that are currently in use. Furthermore, this study addresses the issue of class imbalance within the dataset, a common challenge in skin cancer classification. In addressing this concern, the proposed study leverages the vision transformer and the CNN-based transfer learning models to classify seven distinct types of skin cancers. Through our investigation, we have found that the employment of pre-trained vision transformers achieved an impressive accuracy of 92.14%, surpassing CNN-based transfer learning models across several evaluation metrics for skin cancer diagnosis.

Keywords: skin cancer diagnosis; multi-class; vision transformer; pretrained models; fine tuning; transfer learning; data augmentation



Citation: Arshed, M.A.; Mumtaz, S.; Ibrahim, M.; Ahmed, S.; Tahir, M.; Shafi, M. Multi-Class Skin Cancer Classification Using Vision Transformer Networks and Convolutional Neural Network-Based Pre-Trained Models.

Information **2023**, *14*, 415. <https://doi.org/10.3390/info14070415>

Received: 6 July 2023

Revised: 13 July 2023

Accepted: 14 July 2023

Published: 18 July 2023



Copyright: © 2023 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

One of the most frequent causes of death across the world is cancer. According to the World Health Organization (WHO), approximately 10 million deaths in 2020 were reported to be due to different cancer diseases [1]. Cancer is a condition in which abnormal body cells reproduce uncontrollably and can also spread to other body parts [2]. It is categorized into different types, like lung cancer, breast cancer, which is most common in women, skin cancer, liver cancer, and many more that are the leading causes of human death. Skin cancer is the most common and rapidly spreading cancer that can also cause death. Skin is the

actual physical organ of our body covering different body parts, like the muscles and bones. If the skin is affected by anything, then it has a severe impact on the entire system. Viruses, allergies, alcohol usage, infections, physical activity, exposure to ultraviolet (UV) light, environmental changes, and unusual swellings of the human body are several examples that underly causes of skin cancer. The diseased spot on the skin is called a lesion area, and many lesions are split according to their origin.

Skin cancer is divided into basal cell carcinoma and squamous cell carcinoma. They are referenced as keratinocyte cancers, and their major cause is sun exposure. They frequently affect the body parts directly hit by the sun, like the face, arms, and hands. Another aspect of the body is hardly affected by BCC (basal cell carcinoma), but nearby organs or lymph nodes are easily affected by SCC (squamous cell carcinoma). The other deathly and rare form of skin cancer is melanoma [3], which develops in the melanocytes. It is treatable by surgery if it is detected at the initial stage; otherwise, survival will be difficult or even impossible. Melanoma is a form of cancer that mostly affects white people, often affecting the trunk in males and the lower limb in women; however, it can also appear in other body areas. In the United States, 75% of skin cancer deaths were reportedly caused by melanoma, which constitutes 5% of all skin cancers.

The WHO predicts that skin cancer will be detected in every third cancer patient, so the first and foremost objective of medical research today is to cure skin cancer. This particularly important in this context because as per the statistics, more than twelve million people are reportedly suffering from cancer. The US Skin Cancer Foundation reported that more Americans are affected by skin cancer each year compared to all other cancers; about 5.4 million skin cancer cases are expected to be diagnosed yearly in the US. Therefore, the need for clinical screenings is rapidly increasing. Furthermore, histopathologists find it very difficult to diagnose skin cancer from the epiluminescopy of skin lesions. Doctors generally use the biopsy method to diagnose cancer diseases. This method will assess a different skin sample in the laboratory, which is why this is an excruciating and time-taken procedure.

Macroscopic images generally attained with a digital camera or video are typically analyzed computationally. It is challenging to examine skin lesions if the clinical pictures display the existence of objects, including skin lines, shadows, and hair, in the images and have a poor image resolution. With the help of AI, the morbidity and death burden rate associated with the disease can quickly reduce due to the early diagnosis of diseases. It uses a different technique like machine learning (ML), which contains different algorithms and models that learn from training data. It tries to predict that the output on testing data/new samples to perform the desired tasks is difficult for the human brain. Several types of computer-based systems have been introduced to detect skin cancer. Traditional computer vision algorithms are mostly used as a classifier to diagnose cancer and extract different features from the images. Deep neural networks (DNN), convolutional neural networks (CNN), long short-term memory (LSTM), and recurrent neural networks (RNN) are the most widely used deep-learning models in the healthcare industry.

Traditional ML models need to extract efficient features from the skin images, and skin images are classified on behalf of these valuable features [4]. Due to the feature-length restrictions, ML models are used for skin cancers rather than as a generalized model for different types of diseases [5]. Deep learning (DL) has recently been used for skin cancer classification without an in-depth knowledge of DL and feature extraction. Compared to the ML models, DL models perform well for large-scale datasets [6]. Compared to a dermatologist for cancer identification, ML-based systems are now being designed [7,8], but improved techniques are still required for effective healthcare systems. During the design of the DL model, dataset balancing and many images need to be considered for the practical training and evaluation of the model. Furthermore, DL requires additional training time and costs when the dataset consists of high-resolution images [9]. Kumar et al. [10] proposed a method for evaluating skin infections with a combination of computer vision and a ML approach, from which computer vision was used for feature extraction, and ML

was used for disease evaluation, respectively. The accuracy of the proposed model was 95%. Region-based CNNs have been used for detecting infection with three technologies (region proposal, vector transformation, and classification) in the study published by the authors of [11]. For the classification of skin cancer and other related diseases, the GoogleNet V3 CNN has been used by researchers. They have considered datasets consisting of dermatoscopic images and clinical images of skin cancers and have been able to achieve an accuracy of 72.1 ± 0.9 . Skin lesion classifications and 7-Points checklist techniques have also been used for skin disease diagnosis [12].

The feature extraction method effectively reduces the learning time and improves the performance of the ML and DL models [13]. For the binary classification of clinical images, a CNN-based approach was employed in 2018 [14]. The transfer learning (TL) approach has also been used for the prediction of skin cancer and to achieve the highest accuracy [15]. A novel CNN-based model named SkNet encompassing 19 layers was proposed to classify four types of skin cancer [16]. For skin cancer detection another CNN-based VGG-16 model has also been used for its effective detection performance [17].

In order to recognize and classify skin cancer, early skin cancer is classified by extracting picture attributes, such shape, texture, geometry, and other manually-created methods. CNNs have become the popular method for identifying medical images. CNNs have been successfully used in the categorization of skin cancer due to their remarkable accuracy, demonstrating their value in this field. Although it is possible to extract characteristics for many small objects in an image when using a CNN with a deep architecture, it may be difficult to precisely identify the actually crucial regions. To mitigate this problem, we have employed the vision transformer (ViT) model in this study. The input image is divided into blocks during the general training process for this model, and each block is treated as a separate entity. Self-attention modules are used by the ViT model to understand how these embedded patches are related to one another. The ViT model has recently demonstrated an outstanding performance in the typical classification tasks. The self-attention mechanism of the transformer increases the significance of the important features while reducing the impact of the features that cause noise. Motivated by this perspective, the current study proposed a skin cancer classification network based on the ViT. The findings revealed that the proposed network delivers satisfactory outcomes in skin cancer classification. Furthermore, this study investigated the utilization of CNN-based approaches and fine-tuned them to demonstrate the robustness of the ViT model. This research contributes to the field in the following ways:

- To address the issue of class imbalance, an effective data augmentation technique was implemented to artificially increase the dataset samples;
- The proposed fine-tuned ViT model outperformed the state-of-the-art models for multi-class skin cancer classification;
- In this study, we have also fine-tuned the CNN-based pretrained models, including ResNet50, ResNet101, ResNet152, ResNet50V2, ResNet101V2, ResNet152V2, DenseNet121, DenseNet169, DenseNet201, VGG16, and VGG19, respectively;
- The extensive experiments were performed using the data augmentation technique to propose an effective model;
- The system for classifying multi-class skin lesions has evolved, offering professionals and patients accurate diagnostic information.

2. Materials and Methods

Within this section, we have introduced the methodologies that were adopted and proposed to achieve an effective classification of skin cancer.

2.1. Dataset

For our experiment, we utilized a dataset sourced from Kaggle [18], as shown in Figure 1. However, it is important to note that the dataset we used was imbalanced, which can lead to the presentation of challenges, such as overfitting or underfitting. To ensure an

optimal model performance, it is therefore necessary to address the issue of data imbalances, particularly during the training phase.

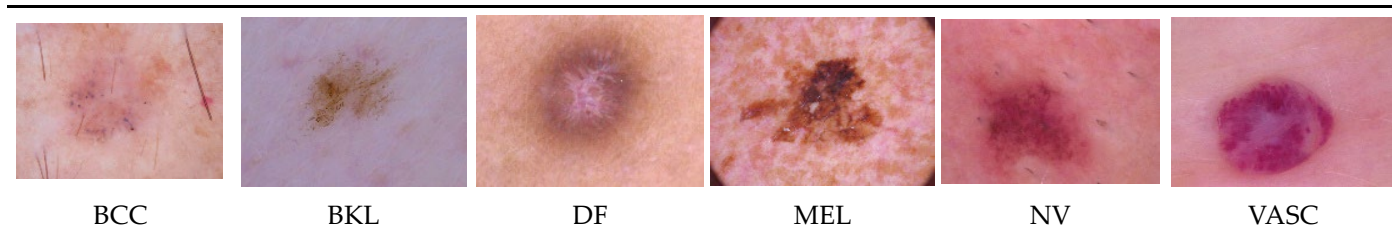


Figure 1. Skin cancer sample images obtained from the HAM10000 dataset.

2.2. Data Augmentation

Data augmentation is a technique that is used to increase the size of a dataset by applying alterations to the original data. It is particularly valuable in deep learning-based models, where having an extensive training dataset is crucial for achieving an effective performance. While data augmentation can be applied across various fields, it is widely used for computer vision problems. In this study, we focused on rotation as an augmentation technique, specifically applying the maximum left and right rotations of up to 8 degrees. By incorporating augmentation, we aimed to diversify the dataset and prevent the introduction of bias in the model caused by class imbalance.

The augmentation process involved two main categories: position and color augmentation. For position augmentation, we utilized sub-techniques such as scaling, cropping, affine transformation, padding, flipping, translation, and rotation. As for color augmentation, we considered techniques like the hue, brightness, saturation, and contrast adjustments. The augmented samples are displayed in Figure 2.

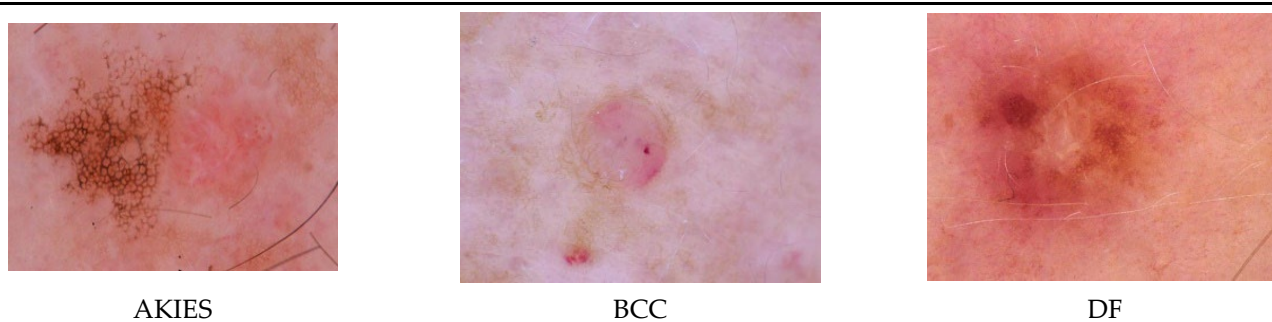


Figure 2. Augmented sample images of the minority classes.

To create a balanced training dataset, we randomly selected and augmented images from the original dataset, ultimately resulting in exactly 980 samples per class just for training. These augmented images were then combined with the original data, resulting in a total of 980 images per class for training (total training samples: 980×7). For validation, we used 140 separate images, and for testing, a separate set of 140 images were considered for this study.

Overall, these techniques, namely data augmentation and random under-sampling, were utilized to address the data imbalance issue and ensure more effective results in our model.

The pre-trained models (e.g., ResNet50, VGG16, and VGG19) were trained and evaluated on this augmented dataset, allowing us to effectively address the class imbalance and enhance the overall performance of the skin cancer classification task.

2.3. Effective CNN-Based Pretrained Model (Fine Tuning of Resnet50)

To ensure effective generalization, we employed minimal preprocessing steps for the proposed model. The dataset images were standardized to a fixed size of 224×224 pixels. Different neural network architectures serve different purposes. For instance, RNNs excel in natural language processing and speech recognition, while CNNs are highly effective in analyzing the visual inputs and processing images.

However, increasing the depth of these networks introduces challenges. Deep networks can be difficult to train due to the vanishing gradient problem, where gradients diminish as they propagate backwards through the layers. To address this issue, a residual neural network (ResNet) was introduced, which employs skip connections. By skipping several layers, the values do not descend to the lowest point, thereby mitigating the vanishing gradient problem. Skip connections involve adding the input to the output of a specific layer [19].

ResNet50, a deep convolutional network comprising 50 layers, was specifically designed for image classification tasks. It was introduced by Kaiming He et al. [20]. ResNet50 consists of two types of blocks: identity blocks and convolutional blocks. Identity blocks are utilized when the input and output dimensions remain the same, while convolutional blocks are used when the input and output dimensions differ. In cases where these dimensions do not match, a convolutional block can be added to the shortcut path to ensure equality between the input and output dimensions.

By incorporating ResNet50 into our research, we aimed to leverage its architecture and skip connections to enhance the performance of our proposed model for skin cancer classification. In this study, the dropout was primarily applied to the hidden and input layers of the neural network, rather than the output layer. This is because during testing, all neurons and connections in the network must be available for accurate predictions and inference. Therefore, the dropout was not applied to the output layer to ensure the availability of the complete network's structure during testing [21]. Further, by utilizing average pooling in our research, we aimed to downsize the feature maps effectively while retaining meaningful information for subsequent layers in the neural network. This allows us to extract the relevant features and facilitate the classification of skin cancer with an improved accuracy and efficiency.

Overfitting is a prevalent challenge encountered in machine learning, where a model performs well on the training data but exhibits unstable predictions on the test data. To address this issue, regularization techniques were employed. One such technique is batch normalization, which plays a vital role in accelerating the training process of the convolutional neural networks and enhancing their stability [22].

Typically, a batch of input data is collected and utilized to train a neural network [22]. By incorporating batch normalization into the training process, we can normalize the activations within each batch, thereby reducing the internal covariate shift. This normalization aids in stabilizing the learning process, allowing the model to generalize better and make more consistent predictions across different datasets.

The activation function is a crucial component in neural networks, determining the activation level of each neuron. It involves calculating the weighted sum of each neuron's input, adding the bias term, and passing the resulting value through the activation function. By introducing non-linearity, the activation function enables neural networks to solve complex problems, as without it, the network would perform linear regression [22].

The activation function operates within specific output ranges, typically limiting values between 1 and -1 , or 0 and 1, respectively. There are two main categories of activation functions: linear/identity activation function and non-linear activation functions. Non-linear activation functions further include various types, such as sigmoid/logistic, SoftMax, Tanh hyperbolic tangent, Leaky ReLU, and ReLU (rectified linear unit) [23,24].

The activation function adds non-linearity to neural networks, allowing them to handle complex tasks effectively. It determines which neurons are active and categorizes the activation functions into linear and non-linear types, each with their specific characteristics

and use cases. Adam [25] is an extension of the SGD (stochastic gradient descent: aims to find the minimum of a given loss function by iteratively adjusting the model's parameters) which is used to recurrently update the network weights based on the training data. This optimizer combines two methodologies—the adaptive gradient algorithm and root mean square propagation—to help sparse gradients on noisy problems. We have considered Adam in this study, as it is better, fast, has low memory requirements compared to other optimization algorithms and has only a few tuning parameters.

Hyperparameter Tuning

Hyperparameters are essential parameters that need to be defined prior to training a model. They play a critical role in determining the behavior and performance of the model. Examples of hyperparameters include the dropout rate, activation function, hidden layer neurons, number of epochs, and batch size [22].

In our study, we empirically determined the values for these hyperparameters, ensuring they are optimized for the specific task at hand. The specific values chosen for each hyperparameter can be found in Table 1, providing transparency and reproducibility in our experimental setup. By carefully selecting and fine-tuning these hyperparameters, we aimed to maximize the performance and effectiveness of our model in classifying skin cancer.

Table 1. Network Hyperparameter Tuning.

Parameter	Values
Hidden neurons	1024, 512, 256, 128, 64, 7
Epochs	20
Dropout	0.3
Activation function	ReLU, SoftMax
Loss-Function	Categorical
Optimizer	Adam
Learning rate	0.001
Batch size	32
Early stopping	Yes
Patience	3

2.4. Transfer Learning and Network Architecture Modifications of the CNN-Based Pretrained Models

Transfer learning is an effective concept that utilizes knowledge acquired from tasks to solve related tasks through fine-tuning. There are several approaches that can be to perform fine-tuning. One approach is to fine-tune some or all of the parameters of the last layer of the pre-trained model [26].

In this work, an integrated feature extractor was utilized to perform effective feature extraction using the pre-trained model. This approach leverages the transfer learning concept, allowing a model trained for a specific problem to be utilized for a different problem by fine-tuning the model. For instance, the last convolutional layer of ResNet50 was mapped with the dense layers (1024, 512, 256, 128, 64, and 7). Batch normalization and a dropout rate of 0.3 were applied after each dense layer. The hidden layers utilized the ReLU activation function, while the last layer consisted of 7 neurons with the SoftMax activation function.

The layers of ResNet50 were frozen, and the weights from the ImageNet dataset were employed in this study. The same configuration was applied to all pre-trained model architectures to maintain consistency.

Overall, both fine-tuning and feature extraction techniques were implemented in this study, with an integrated feature extractor used for effective feature extraction. The specific configurations and approaches described above were applied consistently across all the pre-trained model architectures utilized in our research.

2.5. Vision Transformer (ViT) Pretrained Model

ViT, a deep neural network [27], was designed for image recognition tasks by processing input images through a series of learned transformations. In contrast with the traditional CNNs, ViT employs a self-attention mechanism to focus on the relevant parts of the input image, resulting in high accuracy across the various image recognition tasks.

In this study, the input images were divided into 16×16 patches after resizing the images to 224×224 pixels. The 16×16 patches refer to the process of dividing an input image into smaller fixed-size patches, where each patch was 16 pixels wide and 16 pixels tall. The model was trained on ImageNet-21k, a large-scale image classification dataset with over 14 million images divided into 21,841 categories. This model is composed of 12 transformer layers, each with 768 hidden units and 85.8 million trainable parameters. For the parameter values and configurations of the ViT model, see Table 2.

Table 2. Vision transformer (ViT) configurations.

Parameter	Values
Encoder and pooling layers dimensionality	768
Transformer encoder hidden layers	12
Feed-forward layer dimensionality	3072
Hidden layer activation	Gelu
Hidden layer dropout	0.1
Image size	224×224
Channels	3
Patches	16×16
Balanced	True

Figure 3 presents the abstract-level diagram illustrating the proposed methodology.

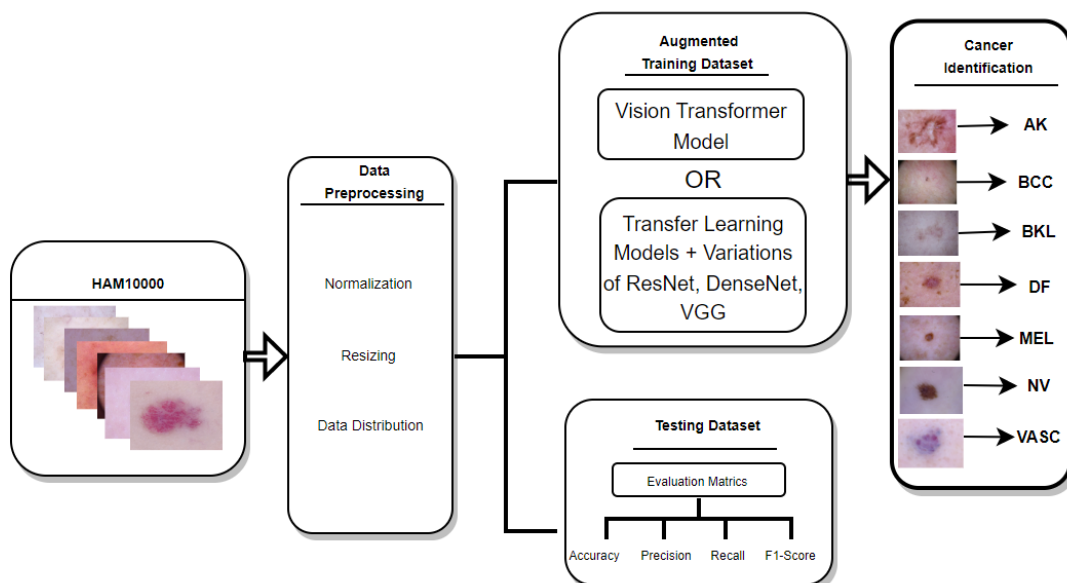


Figure 3. Abstract level diagram of the proposed methodology.

3. Results and Discussion

This section provides a comprehensive discussion on the evaluation measures, experimental details, and the results obtained through the proposed methodology.

3.1. Evaluation Metrics

Evaluation metrics play a crucial role in assessing the performance of the machine learning and deep learning models. These metrics hold significant importance in the realm of machine learning, deep learning, and statistical research. In this study, we have focused on several key evaluation metrics to gauge the effectiveness of our proposed model:

- Accuracy: measures the overall correctness of the model's predictions. It calculates the ratio of correctly classified samples to the total number of samples. Accuracy alone is not always sufficient for evaluation, especially when dealing with imbalanced datasets or when different types of errors have varying consequences;

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

- Precision: quantifies the model's ability to correctly identify the positive **samples** among the predicted positives. It calculates the ratio of true positives to the sum of true positives and false positives. Precision focuses on the reliability of positive predictions;

$$P = \frac{TP}{TP + FP} \quad (2)$$

- Recall: also known as sensitivity or the true positive rate, recall measures the model's ability to correctly identify the positive samples among all actual positives. It calculates the ratio of true positives to the sum of true positives and false negatives. Recall focuses on the completeness of positive predictions;

$$R = \frac{TP}{TP + FN} \quad (3)$$

- F1 Score: the harmonic mean of precision and recall. It provides a single metric that balances both precision and recall, making it useful for when there is an uneven class distribution or an equal emphasis on both types of errors. The F1 score ranges from 0 to 1, with 1 being the best performance.

$$F1 = \frac{(2 \times P \times R)}{(P + R)} \quad (4)$$

Accuracy in terms of multi-class classification is calculated as the ratio of correct predictions (true positives and true negatives) to the total number of predictions, regardless of the class. In comparison, precision, recall, and F1 are considered in the form of weighted averages for multi-class classification. Weighted averaging gives each class a weight based on its proportion in the dataset. To obtain weighted metrics, the precision, recall, and F1 score for each class are multiplied by their corresponding weights, and the results are then totaled and divided by the total weight. This method accounts for the dataset's class imbalance.

3.2. Results and Discussion

The main purpose of this study was to classify skin cancer disease into seven classes. In this study, the model was trained with normal and augmented images. Data augmentation was used in this study to increase the training samples. The model was validated with 140 images, 20 of each class, and was tested with 140 images. The weights transfer technique used the weights of ResNet50 that were retrieved from the ImageNet dataset. The network layers were frozen and tuned with 1024, 512, 256, 128, 64, and 7 neurons of the dense layers, respectively. For the model's generalization, batch normalization and a dropout of 0.3 was

also used in this architecture. It is observable from Figure 4 in that a maximum validation accuracy of 77% and a training accuracy of 86% was achieved with the fine tune ResNet50, respectively.

11-CNN Based Fine Tuned Pretrained Architectures Performance

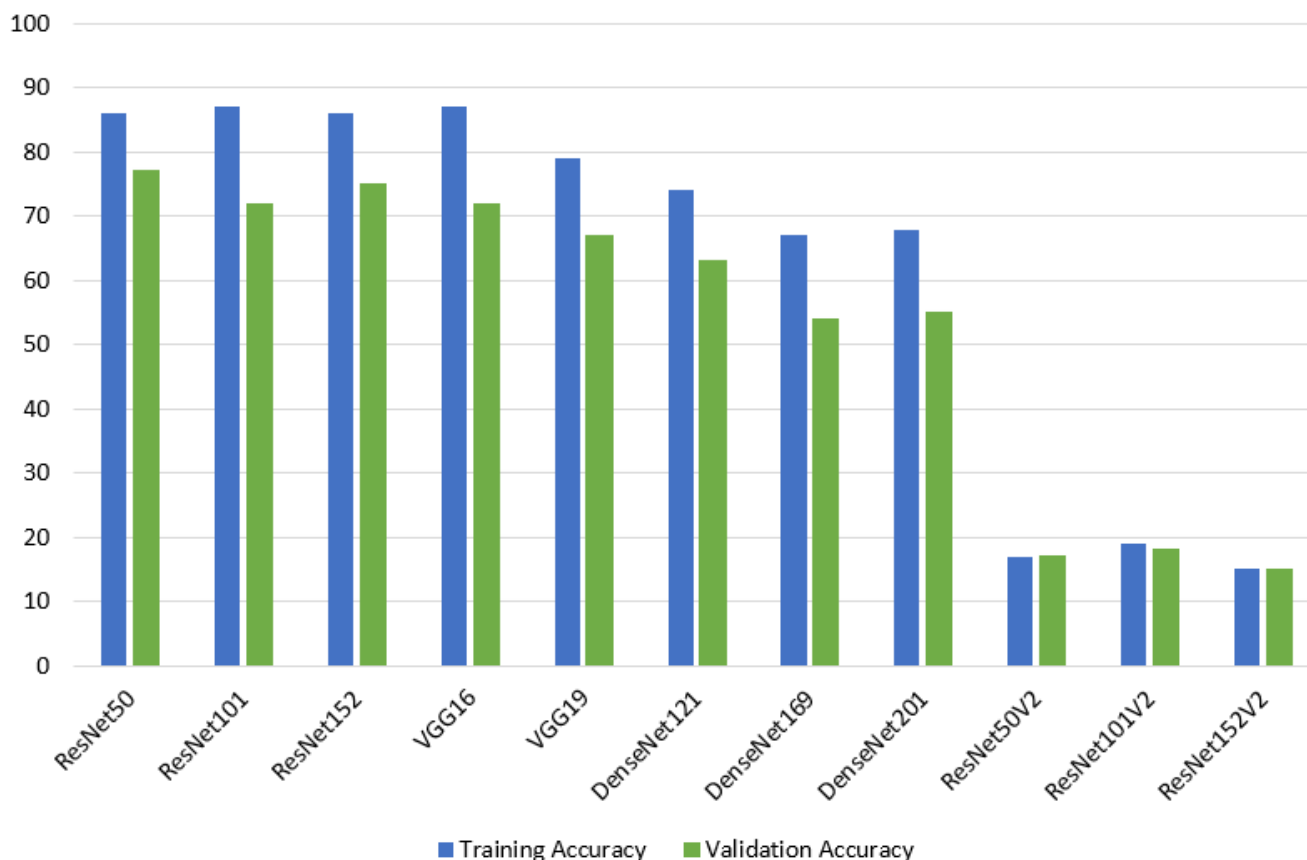


Figure 4. Transfer learning-based fine-tuned pretrained models architectural performances for skin cancer.

We have implemented all the variations of ResNet in this study. The model was also evaluated with test data consisting of 140 images, and we have achieved a test accuracy of 75%. The skin cancer classification problem is a very challenging problem due to the nature of the dataset. The images of the different categories seem the same, raising numerous challenges for the classification model. Although numerous studies have been published in skin cancer classification, the robustness of the proposed model was still deemed to be more accurate. As of now, most of the work has been performed on detecting skin cancer (binary classification).

Sufi A. [28] proposed a study for the classification of melanoma and non-melanoma images and obtained an 83% accuracy. In the study of Hosny K [29], he also proposed the transfer learning approach of the AlexNet model for classifying skin cancer into three classes. The proposed AlexNet model classifies the skin cancer images into melanoma, common nevus, and atypical nevus classes with ~97% accuracy. Dorj U [30] also proposed a study classifying skin cancers into four classes with a 95% accuracy score. The authors of [30] used the ISIC dataset for the skin cancer classification with only two classes: melanoma and non-melanoma.

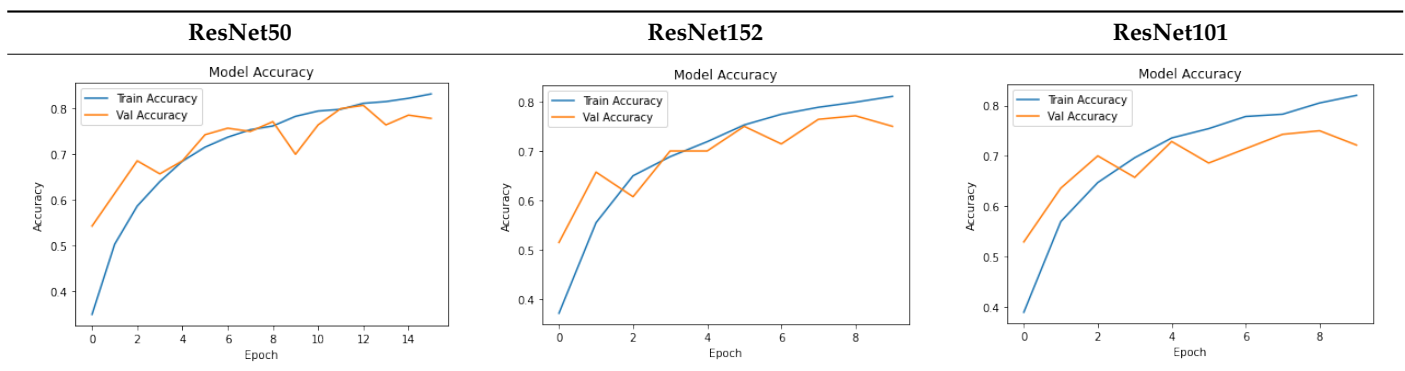
From the analysis of different studies in the literature available on this field, it can be learned that the comparative analysis of skin cancer classification studies is also a challenging task. The dataset varies from study to study, which does not give a fair

comparison. Moreover, the number of classes across the published works varies very frequently. Comparing studies with different datasets and the number of classes will not provide an accurate comparative analysis.

In the study of Budhiman A [31], binary classification was performed on the ISIC dataset and a 0.83 accuracy score was obtained accompanied with a very low positive score (0.46). At the same time, the transfer learning-based ResNet50 fine-tuned model demonstrated a 75% accuracy on the seven types of skin cancer classes in our study. As the increase of values in the target variable increased, the accuracy of the trained model decreased. But the number of classes also increased many times in the proposed study and still obtained a reasonable accuracy score. Hence, the proposed ResNet50 fine-tuned model is a sound addition for classifying the seven types of skin cancers with respect to the CNN-based pretrained model. The performance of all 11 CNN-based pretrained models with same configurations can be seen in Figure 4.

The learning graphs depicting the validation and training processes are essential for gaining insights into the performances of deep learning models during training. These graphs enable us to monitor and analyze the model’s behavior, detect potential overfitting issues, make necessary adjustments to the model’s architecture and hyperparameters, and ultimately enhance the reliability and accuracy of the deep learning models. For a visual representation of the learning graphs, please refer to Table 3, which showcases the learning graphs for the top three pretrained models (ResNet50, ResNet101, and ResNet152).

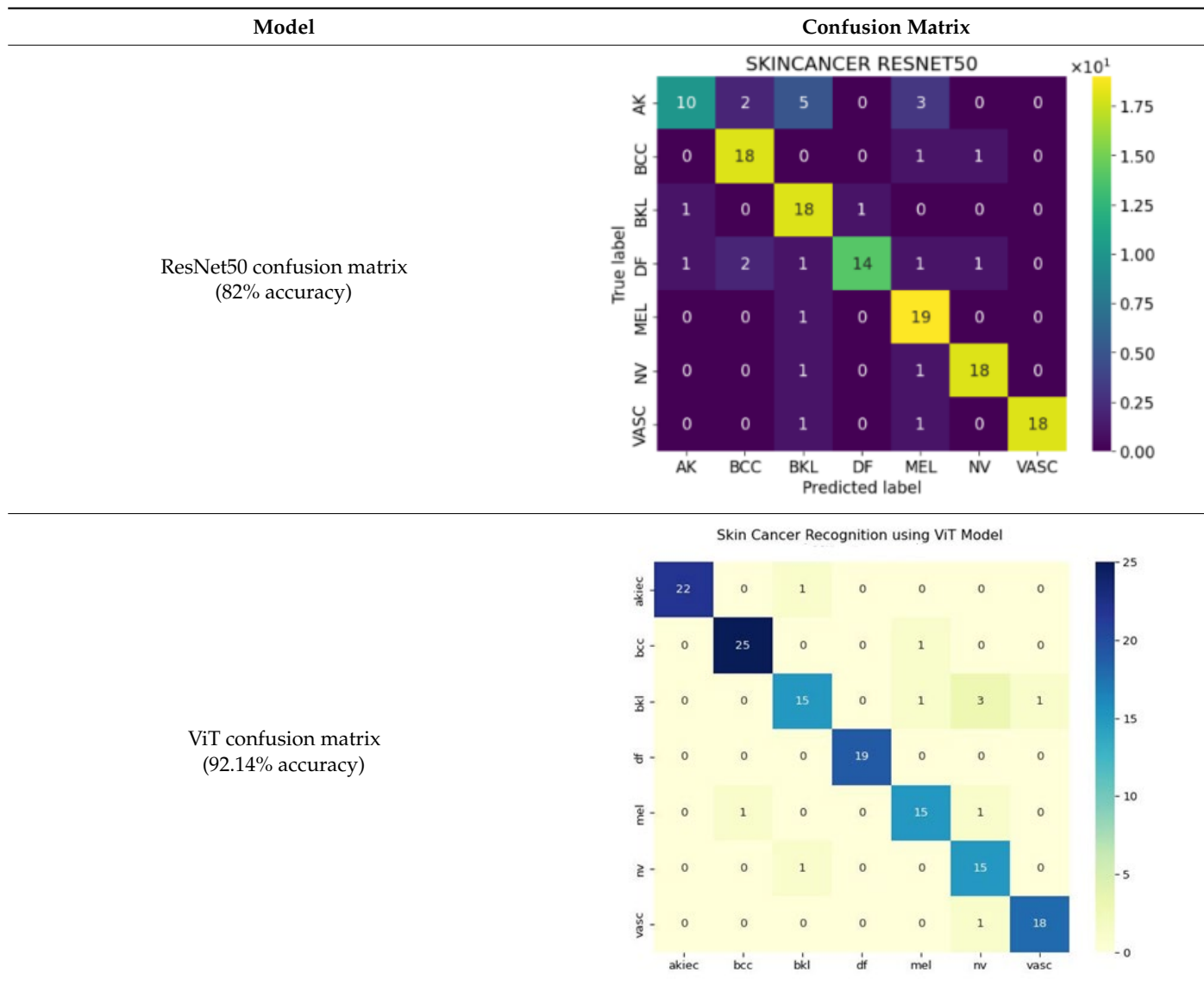
Table 3. Learning graphs of the top three CNN-based pretrained models.



The performance of ResNet50 is superior in terms of its validation accuracy compared to the other conventional pre-trained models. Resnet50 uses a series of convolutional layers to extract features from images before classifying them using fully connected layers, while ViT processes images using a self-attention mechanism. The image is divided into patches, and each patch is given a set of transformer blocks to be applied to it. Due to the transformer blocks and self-attention, the ViT model is more powerful in classifying images. We have considered the balanced dataset with a test ratio of 0.2 for the ViT model and achieved an accuracy of 92.14%, as well as 92.61, 92.14, and 92.17 scores of precision, recall, and F1, respectively.

When there are imbalanced classes or when the cost of misclassifying one class is significantly higher than the cost of misclassifying another class, a confusion matrix can aid in assessing the performance of a classification model. A confusion matrix can derive the accuracy, precision, recall, and F1 score metrics. The comparison confusion matrix of ResNet50 and the ViT model is displayed in Table 4.

Table 4. Confusion matrix of ResNet50 vs the vision transformer model.



The class-wise precision, recall, and F1 score were also used to validate the proposed model, as shown in Table 5. The support column denotes the class sample counts e.g., the AKIEC class has 23 samples for testing while the BCC class has 26 samples for testing purposes. The sum of the support column is equal to 140 as we have tested the model for 140 samples.

Table 5. ViT performance as class wise.

Class name	Precision	Recall	F1	Support
AKIEC	1.0000	0.9565	0.9778	23
BCC	0.9615	0.9615	0.9615	26
BKL	0.8824	0.7500	0.8108	20
DF	1.0000	1.0000	1.0000	19
MEL	0.8824	0.8824	0.8824	17
NV	0.7500	0.9375	0.8333	16
VASC	0.9474	0.9474	0.9474	19

The proposed method outperforms state-of-the-art methods, according to a comparison of the proposed model's performance, as shown in Table 6.

Table 6. Comparison of the proposed study with the state-of-the-art studies.

Authors and Year	Classes	Method	Evaluation Metric	Results
(Ali et al., 2021) [32]	2	Custom CNN-Based Model named DCNN Proposed	Accuracy	Train 93.16% Test 91.93%
(Bassel et al., 2022) [14]	2	Stacking-CV (Proposed) + Xception Features	Accuracy	Test 90.9%
(Jain et al., 2021) [33]	7	Xception Net Transfer Learning-Based Model	Accuracy	Test 90.48%
(Ali et al., 2021) [34]	7	Efficient-Nets B0-B7 Transfer Learning-Based Models (Top Accuracy achieved with Efficient-Net B4)	Accuracy Precision Recall F1	87.91% 88% 88% 87%
(Huang et al., 2023) [35]	3	YoloV5 (RGB Images, HSI Images)	Accuracy Precision Recall F1 Specificity	<u>RGB</u> 79.2% 88.8% 75.8% 81.8% 79.8% <u>HSI</u> 78.7% 80% 72.6% 76.1% 78.6%
Proposed	7	Vision Transformers (RGB Images)	Accuracy Precision Recall F1	92.14% 92.61% 92.14% 92.17%

4. Conclusions

Skin cancer classification is a challenging task due to the diverse appearances of different categories. In this study, we explored the effectiveness of the vision transformer (ViT) approach and pretrained CNN models for multi-class skin cancer classification. By utilizing fine-tuning, transfer learning, and data augmentation techniques, we achieved impressive results. The ViT model outperformed the CNN-based transfer learning models with an accuracy of 92.14%, a precision of 92.61%, a recall of 92.14% and an F1 score of 92.17%, respectively. However, further improvements are needed in preprocessing techniques for deep learning models. Overall, our study highlights the potential of the ViT approach and pretrained CNN models for reliable skin cancer classification.

Author Contributions: Conceptualization, M.A.A., S.M. and M.I.; methodology, S.M., M.A.A., M.T. and S.A.; validation, S.A., M.T., M.S. and M.A.A.; investigation, S.A., S.M. and M.A.A.; data curation, M.A.A. and M.I.; writing—original draft preparation, M.A.A. and S.A.; writing—review and editing, S.A., M.A.A., M.S., M.T. and M.I. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The dataset HAM10000 used in this study for experiments is openly available to download from Kaggle.

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

Abbreviations

AKIEC AK	Actinic Keratoses
BCC	Basal cell carcinoma
BKL	Benign keratosis-like lesions
DF	Dermatofibroma
MEL	Melanoma
NV	Melanocytic nevi
TP	True positive
FP	False positive
TN	True negative
FN	False negative
ViT	Vision transformer
MAX	Maximum
MIN	Minimum
AVG	Average
HSI	Hyperspectral imaging
RGB	Red, green, and blue (color images)

References

1. Cancer. Available online: <https://www.who.int/news-room/fact-sheets/detail/cancer> (accessed on 4 August 2022).
2. Cancer—NHS. Available online: <https://www.nhs.uk/conditions/cancer/> (accessed on 4 August 2022).
3. Melanoma—The Skin Cancer Foundation. Available online: <https://www.skincancer.org/skin-cancer-information/melanoma/> (accessed on 8 July 2023).
4. Arroyo, J.L.G.; Zapirain, B.G. Automated Detection of Melanoma in Dermoscopic Images. In *Computer Vision Techniques for the Diagnosis of Skin Cancer*; Springer: Berlin/Heidelberg, Germany, 2014; pp. 139–192. [CrossRef]
5. Pomponiu, V.; Nejati, H.; Cheung, N.-M. Deepmole: Deep neural networks for skin mole lesion classification. *Proc. Int. Conf. Image Process.* **2016**, *2016*, 2623–2627. [CrossRef]
6. Guo, Y.; Liu, Y.; Oerlemans, A.; Lao, S.; Wu, S.; Lew, M.S. Deep learning for visual understanding: A review. *Neurocomputing* **2016**, *187*, 27–48. [CrossRef]
7. Li, K.M.; Li, E.C. Skin Lesion Analysis Towards Melanoma Detection via End-to-end Deep Learning of Convolutional Neural Networks. *arXiv* **2018**, arXiv:1807.08332.
8. Li, H.; Pan, Y.; Zhao, J.; Zhang, L. Skin disease diagnosis with deep learning: A review. *Neurocomputing* **2021**, *464*, 364–393. [CrossRef]
9. Goyal, M.; Knackstedt, T.; Yan, S.; Hassanpour, S. Artificial intelligence-based image classification methods for diagnosis of skin cancer: Challenges and opportunities. *Comput. Biol. Med.* **2020**, *127*, 104065. [CrossRef] [PubMed]
10. Kumar, V.B.; Kumar, S.S.; Saboo, V. Dermatological Disease Detection Using Image Processing and Machine Learning. In Proceedings of the 2016 3rd International Conference on Artificial Intelligence and Pattern Recognition (AIPR), Lodz, Poland, 19–21 September 2016; pp. 88–93. [CrossRef]
11. Esteva, A.; Kuprel, B.; Novoa, R.A.; Ko, J.; Swetter, S.M.; Blau, H.M.; Thrun, S. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* **2017**, *542*, 115–118. [CrossRef] [PubMed]
12. Kawahara, J.; Daneshvar, S.; Argenziano, G.; Hamarneh, G. Seven-Point Checklist and Skin Lesion Classification Using Multitask Multimodal Neural Nets. *IEEE J. Biomed. Health Inform.* **2019**, *23*, 538–546. [CrossRef] [PubMed]
13. Chandrashekar, G.; Sahin, F. A survey on feature selection methods. *Comput. Electr. Eng.* **2014**, *40*, 16–28. [CrossRef]
14. Haenssle, H.A.; Fink, C.; Schneiderbauer, R.; Toberer, F.; Buhl, T.; Blum, A.; Kalloo, A.; Hassen, A.B.H.; Thomas, L.; Enk, A.; et al. Man against machine: Diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists. *Ann. Oncol.* **2018**, *29*, 1836–1842. [CrossRef] [PubMed]
15. Bassel, A.; Abdulkareem, A.B.; Alyasseri, Z.A.A.; Sani, N.S.; Mohammed, H.J. Automatic Malignant and Benign Skin Cancer Classification Using a Hybrid Deep Learning Approach. *Diagnostics* **2022**, *12*, 2472. [CrossRef] [PubMed]
16. Jeny, A.A.; Sakib, A.N.M.; Junayed, M.S.; Lima, K.A.; Ahmed, I.; Islam, M.B. SkNet: A Convolutional Neural Networks Based Classification Approach for Skin Cancer Classes. In Proceedings of the ICCIT 2020 23rd International Conference on Computer and Information Technology (ICCIT), Dhaka, Bangladesh, 19–21 December 2020. [CrossRef]
17. Tabrizchi, H.; Parvizpour, S.; Razmara, J. An Improved VGG Model for Skin Cancer Detection. *Neural Process. Lett.* **2022**, 1–18. [CrossRef]
18. Skin Cancer MNIST: HAM10000 | Kaggle. Available online: <https://www.kaggle.com/datasets/kmader/skin-cancer-mnist-ham10000> (accessed on 11 September 2022).
19. Residual Neural Network (ResNet). Available online: <https://iq.opengenus.org/residual-neural-networks/> (accessed on 11 August 2022).

20. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. 2016, pp. 770–778. Available online: <http://image-net.org/challenges/LSVRC/2015/> (accessed on 13 July 2023).
21. Dropout Regularization in Neural Networks: How It Works and When to Use It—Programmatically. Available online: <https://programmatically.com/dropout-regularization-in-neural-networks-how-it-works-and-when-to-use-it/> (accessed on 12 August 2022).
22. What Are Hyperparameters? and How to Tune the Hyperparameters in a Deep Neural Network? | by Pranoy Radhakrishnan | Towards Data Science. Available online: <https://towardsdatascience.com/what-are-hyperparameters-and-how-to-tune-the-hyperparameters-in-a-deep-neural-network-d0604917584a> (accessed on 18 August 2022).
23. Activation Functions in Neural Networks—GeeksforGeeks. Available online: <https://www.geeksforgeeks.org/activation-functions-neural-networks/> (accessed on 18 August 2022).
24. What, Why and Which?? Activation Functions | by Snehal Gharat | Medium. Available online: <https://medium.com/@snaily16/what-why-and-which-activation-functions-b2bf748c0441> (accessed on 18 August 2022).
25. Gentle Introduction to the Adam Optimization Algorithm for Deep Learning. Available online: <https://machinelearningmastery.com/adam-optimization-algorithm-for-deep-learning/> (accessed on 18 August 2022).
26. Long, M.; Cao, Y.; Wang, J.; Jordan, M.I.; Edu, J. Learning Transferable Features with Deep Adaptation Networks. *PMLR* **2015**, *37*, 97–105.
27. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An Image is Worth 16 × 16 Words: Transformers for Image Recognition at Scale. *arXiv* **2020**, arXiv:2010.11929.
28. Sufi, A. Skin Cancer Classification Using Deep Learning. 2022. Available online: <http://dSPACE.uuu.ac.bd/handle/52243/2483> (accessed on 15 September 2022).
29. Hosny, K.M.; Kassem, M.A.; Foad, M.M. Skin Cancer Classification using Deep Learning and Transfer Learning. In Proceedings of the 2018 9th Cairo International Biomedical Engineering Conference (CIBEC), Cairo, Egypt, 20–22 December 2018; pp. 90–93. [[CrossRef](#)]
30. Dorj, U.O.; Lee, K.K.; Choi, J.Y.; Lee, M. The skin cancer classification using deep convolutional neural network. *Multimed. Tools Appl.* **2018**, *77*, 9909–9924. [[CrossRef](#)]
31. Budhiman, A.; Suyanto, S.; Arifianto, A. Melanoma Cancer Classification Using ResNet with Data Augmentation. In Proceedings of the 2019 International Seminar on Research of Information Technology and Intelligent Systems (ISRITI), Yogyakarta, Indonesia, 5–6 December 2019; pp. 17–20. [[CrossRef](#)]
32. Ali, M.S.; Miah, M.S.; Haque, J.; Rahman, M.M.; Islam, M.K. An enhanced technique of skin cancer classification using deep convolutional neural network with transfer learning models. *Mach. Learn. Appl.* **2021**, *5*, 100036. [[CrossRef](#)]
33. Jain, S.; Singhanian, U.; Tripathy, B.; Nasr, E.A.; Aboudaif, M.K.; Kamrani, A.K. Deep Learning-Based Transfer Learning for Classification of Skin Cancer. *Sensors* **2021**, *21*, 8142. [[CrossRef](#)] [[PubMed](#)]
34. Ali, K.; Shaikh, Z.A.; Khan, A.A.; Laghari, A.A. Multiclass skin cancer classification using EfficientNets—A first step towards preventing skin cancer. *Neurosci. Inform.* **2022**, *2*, 100034. [[CrossRef](#)]
35. Huang, H.Y.; Hsiao, Y.P.; Mukundan, A.; Tsao, Y.M.; Chang, W.Y.; Wang, H.C. Classification of Skin Cancer Using Novel Hyperspectral Imaging Engineering via YOLOv5. *J. Clin. Med.* **2023**, *12*, 1134. [[CrossRef](#)] [[PubMed](#)]

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.