Enhancing Skin Cancer Detection with Transfer Learning and Vision Transformers

Istiak Ahmad, Bassma Saleh Alsulami, Fahad Alqurashi
Department of Computer Science-Faculty of Computing and Information Technology,
King Abdulaziz University, Jeddah 21589, Saudi Arabia

***Abstract*—Early and accurate detection of skin cancer is critical for effective treatment. This research aims to enhance skin cancer multi-class classification using transfer learning and Vision Transformers (ViTs), addressing the challenges of imbalanced medical imaging data. We introduced data augmentation techniques to the HAM10000 dataset to enhance the diversity of the training and implemented 13 pre-trained transfer learning models. These included DenseNet (121, 169, and 201), ResNet (50V2, 101V2, and 152V2), VGG (16 and 19), NasNet (mobile and large), InceptionV3, MobileNetV2, and InceptionResNetV2, as well as two Vision Transformer architectures (ViT and deepViT). After fine-tuning these models, DenseNet121 achieved the highest accuracy of 94%, while deepViT reached 92%, highlighting the effectiveness of these approaches in skin cancer detection. Future work will focus on refining these models, exploring hybrid approaches that combine convolutional neural networks and transformers, and expanding the framework to other cancer types to advance automated diagnostic tools in dermatology.**

***Keywords*—Medical imaging; skin cancer; multi-class classification; detection; deep learning; transfer learning; vision transformer**

I. Introduction

Skin cancer [1] typically starts in the skin cells when skin cells are damaged, usually from too much exposure to the sun's ultraviolet (UV) rays. The three main types of skin cancer are basal cell carcinoma, squamous cell carcinoma, and melanoma, where melanoma is the most serious. People with fair skin, a history of sunburns, prolonged sun exposure, a family history of skin cancer, or a weakened immune system are at the highest risk. Symptoms of skin cancer include the appearance of new or unusual skin growths, changes in the size, shape, or colour of moles, and persistent non-healing sores. According to the American Cancer Society [2], in 2024, there are estimated to be 100,640 new cases of melanoma skin cancer (59,170 male and 41,740 female) and an estimated 8,290 deaths (5,430 male and 2,860 female). The National Cancer Institute [3] reported that the relative survival rate for skin cancer over the five years from 2014 to 2020 is 94.1%. Early skin cancer detection is crucial, as it significantly enhances the likelihood of effective therapy and survival.

Artificial intelligence plays a pivotal role in enhancing sustainability across various sectors, including education [4], transportation [5], cybersecurity [6], [7], social media [8], [9] and healthcare. Notably, the field of healthcare has experienced remarkable advancements, particularly in the early detection of cancer, with AI's contributions proving to be increasingly impactful. The prompt identification of skin cancer enables using less invasive treatment approaches, reducing the risk of severe consequences and improving the patient's prognosis. The process of developing an early skin cancer detection system involves integrating cutting-edge imaging technology with advanced deep-learning algorithms. These appliances have the capability to accurately analyse skin lesions, detecting distinguishing patterns that may signify the presence of cancer. Computer-aided diagnosis (CAD) is a vital component of these systems, providing dermatologists with a valuable tool. CAD systems enhance diagnostic accuracy by analyzing visual data from skin images and distinguishing between benign and malignant tumours. CAD systems are essential for improving skin cancer's early and accurate diagnosis [10], improving patient outcomes by providing reliable views, and eliminating diagnostic errors. Skin cancer detection encounters challenges, including limited, diverse datasets, which hinder model generalization and class imbalance, leading to biased predictions. The visual similarity between benign and malignant lesions complicates detection, while the "black box" nature of AI models raises interpretability issues. Overcoming these hurdles is essential for enhancing skin cancer detection technologies. Transfer learning and Vision Transformers (ViTs) enhance model performance and generalization to overcome challenges with skin cancer detection. Transfer learning utilizes pre-trained models on extensive datasets to improve accuracy, even when skin cancer data is limited, and to address the issue of imbalanced classes. Vision Transformers (ViTs) improve the distinction between visually similar benign and malignant tumours by analyzing complex patterns and establishing connections across long distances in images.

A. The Aim and Objectives

The aim of this research is to develop and evaluate a robust multiclass skin cancer detection system by leveraging transfer learning, Vision Transformers, and advanced data augmentation techniques to achieve high accuracy. The research contributions are as follows:

- Applied refined data augmentation techniques to address challenges of limited and imbalanced datasets, significantly improving model robustness and generalization.
- Conducted a comprehensive analysis of 13 transfer learning models, underlining their comparative performance and demonstrating significant improvements in skin cancer detection accuracy.
- Implemented two Vision Transformer models, effectively leveraging their ability to capture intricate patterns in skin lesion images, enhancing detection accuracy.

The remainder of the paper is organised as follows: Section II discusses the literature review and research gaps. Section III discusses the proposed research methodology. Section IV discusses the research outcomes for transfer learning models and vision transformers models. Section V compares the overall research outputs with existing research and discusses research challenges and limitations. Finally, Section VI concludes by stating the future direction.

II. Literature Review

Kondaveeti et al. [11] proposed pre-trained models, such as ResNet50, MobileNet, Xception, and InceptionV3 for 7 types of skin cancer detection. They achieved the highest 90% accuracy, 89% weighted average precision, and 90% recall for ResNet50 on the HAM10000 dataset. Naik et al. [12] proposed MobileNetV2 for skin cancer detection and achieved 93.11% accuracy. Fraiwan and Faouri [13] used 13 transfer learning models to detect 7 types of skin cancer. They achieved the highest 82.9% accuracy for the DenseNet201 model. Swetha et al. [14] compared several transfer learning models, such as ResNet50, ResNet152, ResNet101, VGG16, VGG19, MobileNet, and Xception for multiclass skin cancer detection and obtained 83.69% categorical accuracy. Vishnu et al. [15] presented augmentation techniques and ensemble models by integrating InceptionV3 and DenseNet201 weights and outputs for six types of skin cancer detection. They achieved 89% accuracy with a validation loss of 0.44. Kaveti et al. [16] proposed a ResNet101 model for multiclass skin cancer detection using the HAM1000 dataset and got an accuracy of 92% for seven skin cancer categories. Arshed et al. [17] employed a pre-trained vision transformer model for seven types of skin cancer detection and achieved 92.14% accuracy. Tuncer et al. [18] presented lightweight CNN-based techniques for detecting benign and malignant types of skin cancer. They achieved 92.12% accuracy using the HAM10000 dataset. Yang et al. [19] proposed attention-weighted transformers for skin cancer detection using the HAM10000 dataset and achieved 93.75% accuracy. Ashfaq and Ahmad [20] implemented InceptionResNetV2, VGG16, ResNet50, EfficientNetB3, and vision transformer B32 for multiclass skin cancer classification. Among these models, EfficientNetB3 obtained the highest accuracy of 86% with precision 85%, recall 79%, and F1-score 81%. Sanchez et al. [21] proposed EfficientNet (B0, B1, B2, and B3) and ViT (base 16, base 32, large 16 and large 32) models and obtained the highest accuracy of 85% by fine-tuning the number of epochs and learning rate.

Despite considerable progress in skin cancer detection, key gaps remain that align with this study's objectives. Existing research has explored data augmentation to address imbalanced datasets; however, more refined techniques are needed to further enhance model robustness. This study addresses the scarcity of comprehensive comparisons of multiple transfer learning models' performances by evaluating 13 models, thereby providing a clearer insight into their effectiveness. Moreover, Vision Transformers (ViTs) have shown promise, but their potential to capture detailed patterns in skin lesion images remains underexplored. This research advances the field by implementing and testing two ViT models, demonstrating their capability to enhance detection accuracy.

III. Methodology and Design

Fig. 1 presents the proposed research methodology for developing and evaluating models for cancer detection using transfer learning and vision transformer techniques. The methodology is divided into several key steps. The first step includes dataset collection from the HAM10000 dataset. We have conducted several exploratory data analyses, including analysing the distribution of classes and disease distribution across genders, to acquire an understanding of the dataset. We have discovered that the dataset is imbalanced, which can lead to bias and overfitting issues. Data augmentation techniques are implemented to resolve these issues and improve model generalization. The more details of the dataset distribution are discussed in Section III-A. This study implemented several transfer learning models (see Section III-B) and vision transformer models (see Section III-C) for cancer detection. Finally, the models are evaluated using evaluation metrics (see Section III-D).

Algorithm 1 shows the process of our proposed methodology. Initially, we read all the images from the directory and saved them with the corresponding labels (line numbers 23 and 24). After that, image preprocessing is employed to remove noise from the image, and image augmentation techniques are applied to balance the number of images for each class (lines 25 and 26). The function AUGMENTEDPIPELINE discusses the applied augmentation techniques in this study (line numbers 1 to 11). Following this, the dataset is split into training, validation, and testing sets equally (line number 27). Then, the Keras transfer learning pre-trained model is loaded as a base model and added to a fully connected layer (FCmodel) (line numbers 29 and 30). The detailed architecture of a fully connected layer is given from lines numbers 12 to 22. Furthermore, the FCmodel, Adam optimizer, and entropy loss function are passed as parameters to the compiler (line number 31). Similarly, PyTorch vision transformer modes (DeepViT and ViT) are loaded as transformer models (line number 33). To calculate the training time of the proposed method, a timer is started in line number 34. Sequentially, the compiled model, validation, and training set are passed to train the model (line number 35). After training the model, the model and training history are saved in h5 format (line number 36). Line number 37 provides the model execution time. Finally, we evaluate our proposed method by passing the testing set, saved model, and training history through several evaluation metrics.

A. Dataset

This study used the HAM10000 (Human Against Machine with 10000 training images) [22] dataset, which was presented by the International Skin Imaging Collaboration (ISIC) for skin cancer detection in 2018 [23]. This dataset contains 10,015 dermatoscopic images for seven types of cancer, including Actinic keratoses and intraepithelial carcinoma (akiec), basal cell carcinoma (bcc), benign keratosis-like lesions (bkl), dermatofibroma (df), melanoma (mel), melanocytic nevi (nv) and vascular lesions (vasc). Fig. 2 presents the distribution of seven types of skin cancer derived from the HAM10000 dataset. The image IDs, along with the corresponding patient details, such as age, sex, and lesion localization, are as follows: akiec (ISIC_0027930, 60, male, and scalp), bcc (ISIC_0026343, 70, male, and face), bkl (ISIC_0028233, 55, male, and

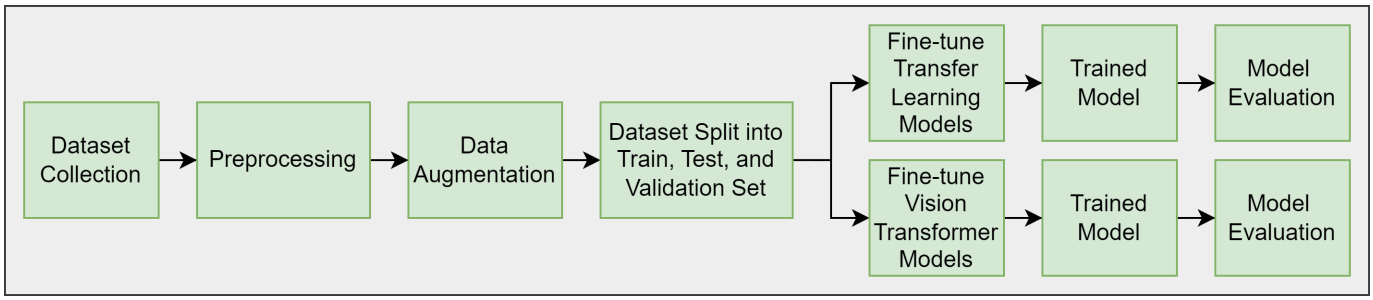

Fig. 1. Research methodology and design.

Algorithm 1 Master Algorithm

Input: *directory path of mages*
Output: *skin cancer classification*

function AUGMENTEDPIPELINE($imgProcess$)
  $imgProcess.Sequential([$
  $imgProcess.Fliplr(0.5),\ \#\ horizontal\ flip$
  $imgProcess.Flipud(0.5),\ \#\ vertical\ flip$
  $imgProcess.Crop(percent = (0,\ 0.1),$
  $imgProcess.MultiplyBrightness(0.8,\ 1.2),$
  $imgProcess.GaussianBlur(sigma = (0, 1.0))),$
  $imgProcess.Grayscale(alpha = (0.0, 1.0)),$
  $imgProcess.CoarseDropout(0.02),$
  $imgProcess.CLAHE(clip\_limit = (1, 4))])$
end function
function FCLAYER($baseModel$)
  $A \leftarrow baseModel.output$
  $A \leftarrow GlobalAveragePooling2D()(A)$
  $A \leftarrow Flatten()(A)$
  $A \leftarrow Dense(128, activation = "relu")(A)$
  $A \leftarrow Dropout(0.3)(A)$
  $A \leftarrow Dense(512, activation = "relu")(A)$
  $A \leftarrow Dropout(0.3)(A)$
  $A \leftarrow Dense(7, activation = "softmax")(A)$
  $FCmodel \leftarrow Model(baseModel.input, A)$
end function
$imageDir \leftarrow directory\ path\ of\ mages$
$imgPathLabel \leftarrow readImages\ (imageDir)$
$imgProcess \leftarrow imagePreprocessing\ (imgPathLabel)$
$imgAugment \leftarrow augmentedPipeline\ (imgProcess)$
$train, test, valid \leftarrow dataset\ split\ (imgAugment)$
$\#\ Transfer\ Learning\ Models:\ using\ Tensorflow, Keras$
$baseModel \leftarrow pretrainedModel(model)$
$FCmodel \leftarrow FClayer(baseModel)$
$Model \leftarrow compileModel(FCmodel, optimizer = adam, loss = entropy)$
$\#\ Transformer\ Models:\ using\ PyTorch$
$Model \leftarrow transformer(baseModel, parameters, optimizer = adam, loss = entropy)$
$start \leftarrow time.time()$
$Fmodel \leftarrow trainModel(Model, train, validation)$
$savedModel, history \leftarrow save(Fmodel)$
$executeTime \leftarrow time.time() - start$
$evalMetrics \leftarrow evaluation(test, savedModel, history)$

face), df (ISIC_0028880, 55, male, and lower extremity), nv (ISIC_0028888, 50, male, and trunk), mel (ISIC_0029241, 70, male, and face), and vasc (ISIC_0027790, 50, female, and face).

Table I provides a comprehensive overview of the HAM10000 dataset, demonstrating the distribution of images across different classes and dataset splits. Each class includes a specific number of original and augmented images, as well as their distribution across training, validation, and test sets. As the original dataset is imbalanced, we implemented data augmentation techniques such as horizontal and vertical flipping, cropping, rotation, adjusting brightness and contrast, applying Gaussian blur, and adding Gaussian noise to address the class imbalance and improve the model's generalization capability.

TABLE I. DISTRIBUTION OF ORIGINAL, AUGMENTED, AND SPLIT DATASET IMAGES ACROSS SEVEN DIAGNOSTIC CLASSES

Classes	Original Images	Augmented Images	Train	Valid	Test
akiec	327	6705	3747	961	1997
bcc	514	6705	3710	958	2037
bkl	1099	6705	3739	923	2043
df	115	6705	3810	898	1997
mel	1113	6705	3801	925	1979
nv	6705	6705	3751	913	2041
vasc	142	6705	3725	993	1987
Total	10015	46935	26283	6571	14081

After applying data augmentation, a uniform count of 6,705 augmented images per class was generated. We split the data into training, validation, and test sets, utilizing 26,283 images for training, 6,571 for validation, and 14,081 for testing. These splits ensure a balanced representation of each class. This detailed distribution of images allows the model to be well-trained and evaluated, with a balanced exposure to each class, thereby reducing the risk of bias and overfitting.

B. Transfer Learning Model

This study used 13 transfer learning models, such as like DenseNet (121, 169, and 201), ResNet (50V2, 101V2, and 152V2), VGG (16 and 19), InceptionV3, MobileNetV2, and NASNet (Mobile and Large). We got the highest accuracy for the DenseNet121 model. The DenseNet121 architecture, as illustrated in Fig. 3, is a robust deep learning model optimized for image classification tasks through its efficient feature extraction capabilities. The model begins with an input layer that processes images of size 224x224 with three colour

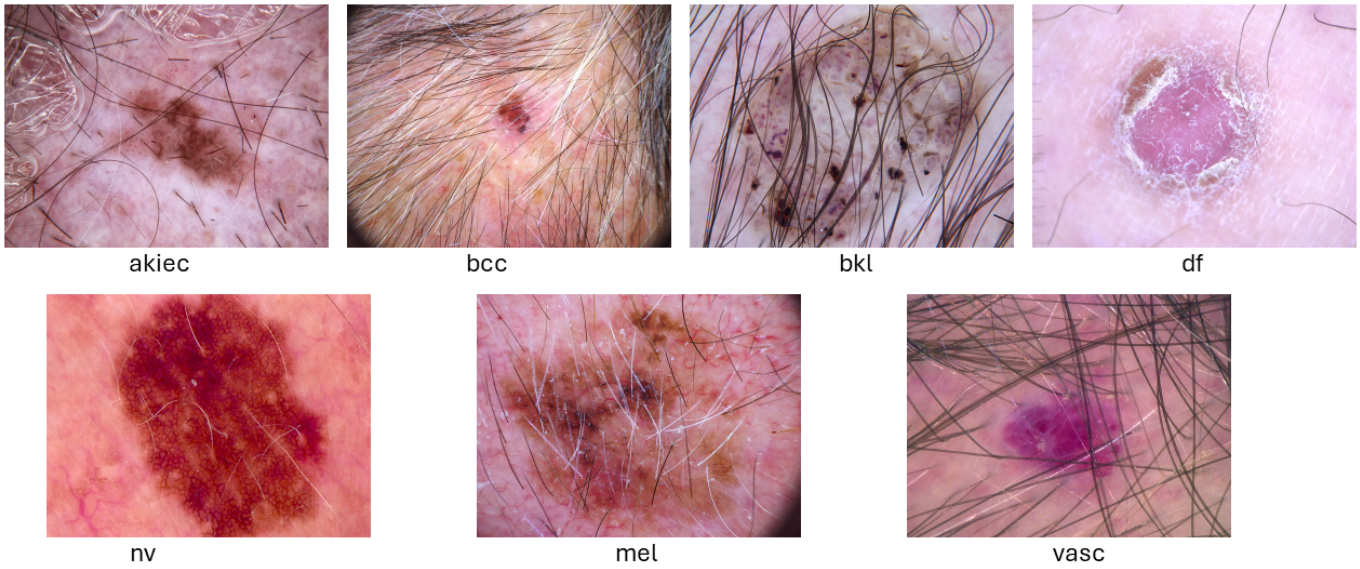

Fig. 2. Seven types of skin cancer.

channels (RGB). The next layer is an initial convolutional layer to capture fundamental features, such as edges. This layer produces a feature map with 64 channels and a spatial dimension of 56x56. The core of the model lies in its dense blocks (D1-D4), where each layer connects to all previous layers within the same block, ensuring that features are reused and information flows efficiently through the network. These dense blocks are interspersed with transition layers (T1-T3), which use 1x1 convolutions and average pooling to reduce the size of the feature maps, thereby maintaining computational efficiency without losing critical information. As the network progresses, the number of channels increases while the spatial dimensions decrease, refining the features extracted. The final layers include a Global Average Pooling (GAP) layer, which condenses the feature map into a vector, followed by fully connected layers that prepare the model for classification. Dropout layers are integrated to prevent overfitting, and the model concludes with a softmax activation function that outputs the probabilities for each class, making DenseNet121 both robust and accurate in image classification tasks.

C. Vision Transformer Model

Fig. 4 compares the architectures of the Vision Transformer (ViT) on the left and DeepViT on the right, highlighting key differences in their approaches to attention mechanisms and processing layers. Both models start with patch embedding and positional embedding to transform input images into a sequence of vectors, enabling the application of transformer-like attention mechanisms. In ViT, the architecture employs a standard multi-head self-attention mechanism followed by normalization and feed-forward layers. This process is repeated multiple times (xN), allowing the model to learn and refine features through self-attention across all patches. The simplicity of ViT lies in its straightforward use of self-attention without any modifications, relying on the depth of the network to capture complex patterns. However, DeepViT presents a modification known as "Re-Attention". After the initial multi-head attention, DeepViT applies normalization and a matrix transformation to refine the attention weights before they are fed into the next layers. This re-attention mechanism aims to stabilize and improve the quality of attention by reinforcing the most critical relationships between patches, potentially leading to better performance in capturing long-range dependencies.

D. Evaluation Metrics

This study used the following metrics to evaluate the performance of the classification model: A True Positive (TP) occurs when a model accurately predicts a positive class, whereas a True Negative (TN) occurs when it correctly predicts a negative class. A False Positive (FP) results when a model erroneously predicts a positive class and a False Negative (FN) happens when a model wrongly predicts a negative class as positive. These metrics were used to calculate the accuracy, precision, recall, and F1-score (see Eq. 1).

$$\begin{aligned} Accuracy &= \frac{TP + TN}{TP + TN + FP + FN} \\ Precision &= \frac{TP}{TP + FP} \\ Recall &= \frac{TP}{TP + FN} \\ F1 - score &= \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} \end{aligned} \tag{1}$$

Each metric is crucial for evaluating the model's effectiveness in classifying the lesions accurately. Accuracy indicates the overall correctness of the model's predictions. Precision measures the model's ability to correctly identify positive cases, while recall assesses how well the model captures all actual positive cases. The F1-Score, a harmonic mean of precision and recall, provides a balanced measure of the model's performance.

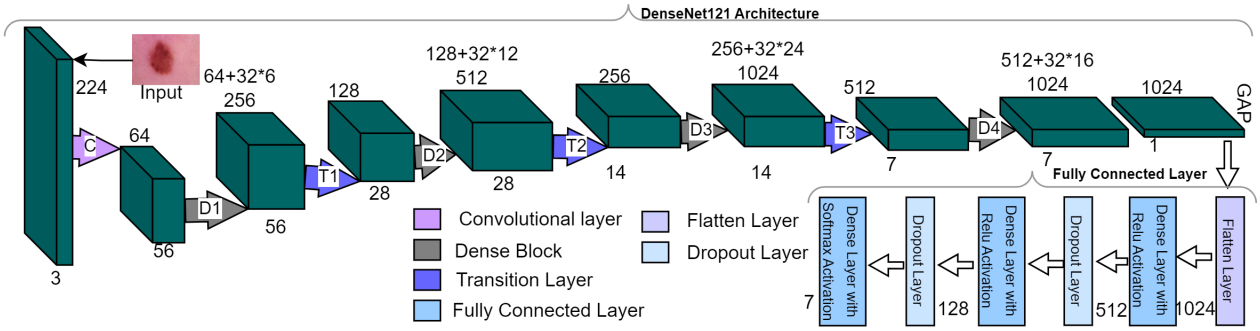

Fig. 3. DenseNet121 architecture.

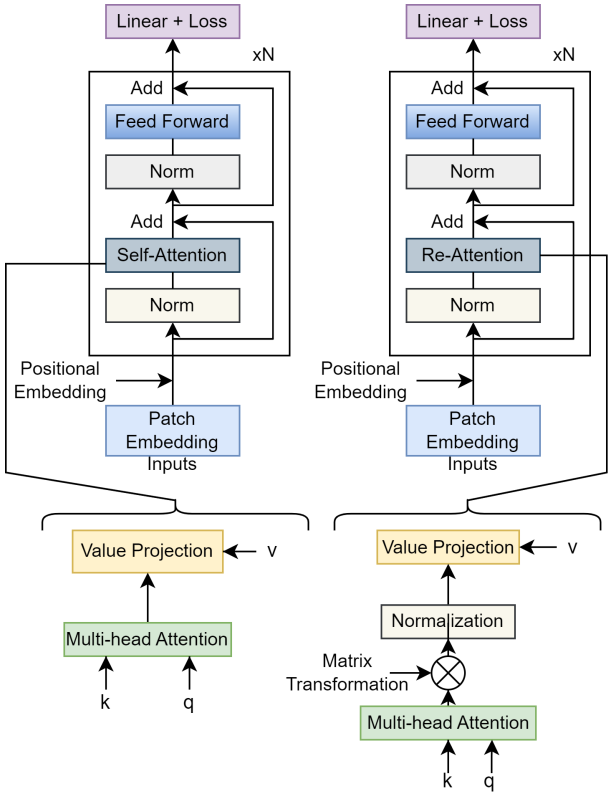

Fig. 4. Comparison between ViT (Left) and DeepViT (Right) model architecture [24].

1) Experimental Setup: The experiment was implemented on the following GPU configurations: NVIDIA GeForce RTX 4090, with 24GB memory. We employed the following parameters to train the transfer learning models: image size (224 x 224), batch size = 32, loss function: "categorical_crossentropy", optimizer: "adam", and two callback functions (ReduceLROnPlateau and EarlyStopping). For the deep vision transformer, we used the following parameters: batch size = 32, epochs = 100, learning rate = 0.00001, gamma = 0.7 and seed = 42, patch size = 32, and embedding dropout = 0.1.

TABLE II. PERFORMANCE COMPARISON OF VARIOUS DEEP LEARNING MODELS1

	Models	Ep	Time (Hour)	Acc	Pre	Re	F1-Score
Transfer Learning	DenseNet121	16	1.94	**0.94**	**0.94**	0.99	**0.94**
	DenseNet169	13	9.87	0.92	0.92	**1.00**	0.92
	DenseNet201	13	10.55	0.92	0.92	**1.00**	0.92
	ResNet50V2	16	3.95	0.92	0.93	0.98	0.93
	ResNet101V2	18	5.46	0.93	0.93	0.99	0.93
	ResNet152V2	26	11.20	0.93	**0.94**	0.99	0.93
	VGG16	12	5.47	0.89	0.91	0.97	0.90
	VGG19	12	5.2	0.87	0.88	0.97	0.87
	InceptionResNetV2	15	3.19	0.91	0.92	0.99	0.91
	InceptionV3	40	7.44	0.90	0.91	0.99	0.90
	MobileNetV2	9	1.49	0.86	0.87	0.98	0.86
	NASNetMobile	30	8.68	0.91	0.91	0.99	0.91
	NASNetLarge	22	8.01	0.90	0.90	0.97	0.90
TF	ViT	80	8.92	0.91	0.92	0.91	0.92
	DeepViT	93	10.56	**0.92**	**0.93**	**0.92**	**0.93**

Ep = Epochs, Acc = Accuracy, Pre = Precision, Re = Recall

IV. RESULT

Table II compares the performance of various deep-learning models employed to classify skin cancers. The models report key performance metrics like accuracy (Acc), precision (Pre), recall (Re), and F1-score, along with the number of epochs (Ep) and training time in hours. For transfer learning, DenseNet121 emerges with the highest accuracy, precision, and F1-score of 0.94 with only 16 epochs and a relatively short training time of 1.94 hours. Additional variations of DenseNet, such as DenseNet169 and DenseNet201, exhibit strong performance while needing longer training durations. ResNet models exhibit impressive performance, especially ResNet152V2, which achieves an optimal trade-off between accuracy (0.93) and precision (0.94) but with the most extended training duration of 11.20 hours over 26 epochs. However, VGG models demonstrate the least accuracy, with VGG16 and VGG19 achieving accuracies of 0.89 and 0.87, respectively. The InceptionV3 and NASNet models demonstrate satisfactory performance, with NASNetLarge and InceptionV3 achieving an accuracy of 0.90. For the vision transformer, the ViT model, while requiring a significantly higher number of epochs (80) and more prolonged training time (8.92 hours), achieves an F1-score of 0.92, indicating robust performance. DeepViT slightly outperforms ViT with a higher F1-score of 0.93, albeit with a longer training time of 10.56 hours.

Fig. 5 illustrates the progression of training and validation loss, as well as accuracy measures, during best 15 epochs for a DenseNet121 model. The training loss, shown by the

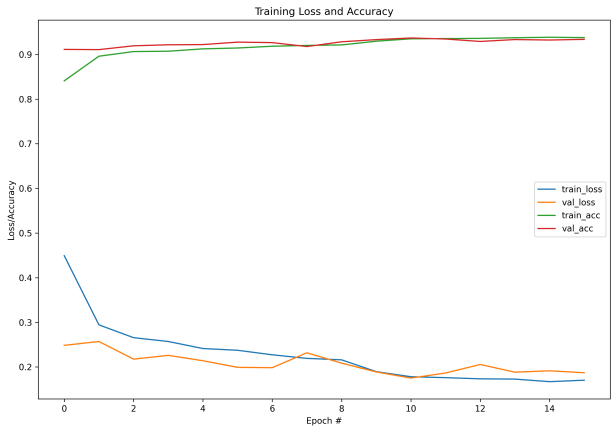

Fig. 5. Training & validation loss and accuracy (DenseNet121).

blue line, initially begins at a very large value and gradually drops, suggesting that the model is progressively learning and improving its performance as time progresses. The validation loss, shown by the orange line, has a similar trend to the training loss but consistently maintains a slightly lower value throughout all epochs. This suggests that the model possesses high generalization capabilities. The training accuracy, shown by the green line, exhibits a rapid initial rise, reaching a plateau by the third epoch, and then maintains a consistently high level for the remaining epochs. The validation accuracy, shown by the red line, has a similar pattern to the training accuracy, indicating that the model consistently performs at a high level on both the training and validation datasets.

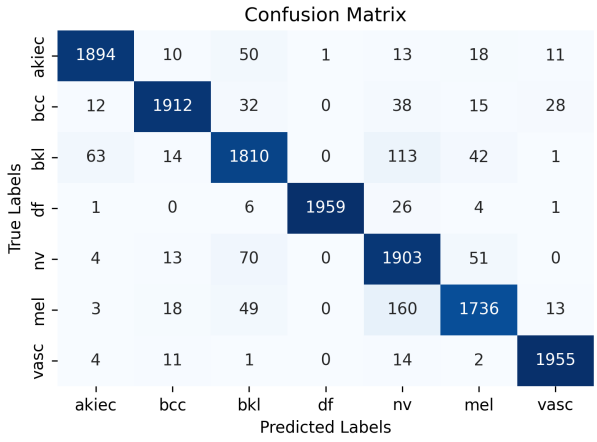

Fig. 6. Confusion matrix (DenseNet121).

Fig. 6 demonstrates the efficacy of a DenseNet121 model on a multi-class skin lesion dataset. The model has high accuracy, accurately categorizing the majority of cases in each category. For example, it accurately recognized 1894 occurrences of akiec, with the majority of misclassifications happening as bkl (50 occurrences). Similarly, the bcc category was accurately identified 1912 times, although some cases of misunderstanding resulted in misclassifications as nv (38 occurrences) and bkl (32 occurrences). The model accurately predicted 1810 cases of bkl. However, it incorrectly categorized some as mel (113) and akiec (63). The df class achieved near-perfect classification, with 1959 accurate predictions and few mistakes. The nv class achieved 1903 accurate classifications, but the model exhibited confusion in distinguishing between mel (70 occurrences) and bcc (51 cases). In the mel class, 1736 occurrences were accurately categorized, whereas 160 were mistakenly classified as nv. Ultimately, the model successfully categorized 1955 occurrences of vasc, with few mistakes. In general, the matrix demonstrates a high level of accuracy in classifying the data; however, there are a few instances where comparable classes were mistakenly identified, namely between mel and nv, as well as between akiec and bkl.

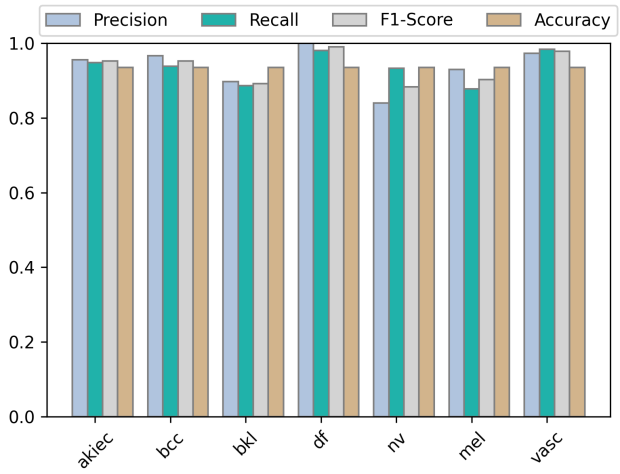

Fig. 7. Precision, recall, F1-Score, accuracy (DenseNet121).

Fig. 7 presents the performance metrics—Precision, Recall, F1-Score, and Accuracy—across different skin lesion classes (akiec, bcc, bkl, df, nv, mel, vasc) for a DenseNet121 model. The figure shows consistently high values across all metrics, demonstrating the model's strong performance across different classes, though slight variations suggest areas for potential improvement in balancing precision and recall for certain classes.

Fig. 8. Training & validation loss and accuracy (DeepViT).

Fig. 8 displays the training and validation loss and accuracy curves over the course of training for a DeepViT model, specifically focusing on the convergence and performance trends across epochs. The training loss (orange line) decreases

steadily, indicating effective learning, while the validation loss (blue line) also decreases but shows more fluctuation, suggesting variability in model performance on unseen data. The training accuracy (red line) improves rapidly and plateaus, reflecting the model's increasing ability to correctly classify training samples. The validation accuracy (green line) follows a similar trend but with smaller fluctuations. This implies that the model generalizes reasonably well but may still be prone to some overfitting, as indicated by the gap between the training and validation losses. Overall, the figure suggests that the model has learned effectively, but the slight divergence between training and validation metrics could indicate room for further tuning to improve generalization.

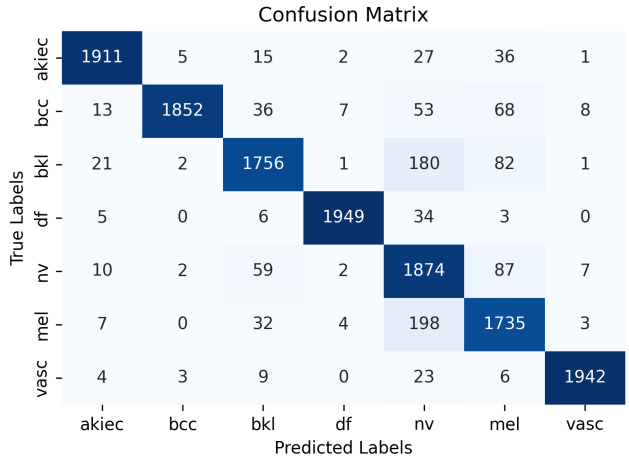

Fig. 9. Confusion matrix (DeepViT).

Fig. 9 illustrates the performance of the DeepViT transformer model in classifying various types of skin cancer. The diagonal elements represent the correctly classified instances for each class, indicating strong model performance, particularly for classes such as akiec, bcc, df, and vasc, with high true positive counts. However, some misclassifications are observed, especially in classes bkl and mel, where the model mistakenly predicts other categories. For example, 180 instances of bkl were misclassified as nv, and 82 instances of mel were also confused with nv.

Fig. 10 presents the performance metrics—Precision, Recall, F1-Score, and Accuracy—of the DeepViT transformer model across different skin cancer classes. The model shows consistently high performance across most classes, with metrics generally above 0.8, indicating effective classification. Notably, the classes akiec, bcc, and vasc exhibit near-perfect scores in all metrics, reflecting the model's strong ability to distinguish these types. These results demonstrate the potential of the DeepViT transformer in enhancing the accuracy of automated skin cancer detection, although further improvements are necessary to address the specific misclassification challenges observed in certain classes.

V. Discussion

Table III presents a comparison of the accuracy achieved by different deep-learning models for skin cancer detection. All the presented articles used the same dataset for multiclass skin

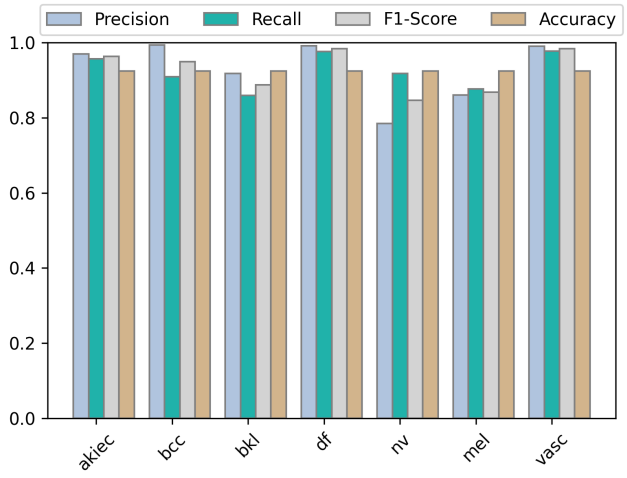

Fig. 10. Precision, recall, F1-Score, accuracy (DeepViT).

cancer classification. It places the DenseNet121 model from the current study in context with other existing methodologies, such as ResNet50, MobileNetV2, DenseNet201, Vision Transformers (ViT), EfficientNet, and others. The reported accuracy ranges from 82.9% to 93.75%. The DenseNet121 model from this study notably achieved an impressive accuracy of 94%, surpassing all the other models listed. This outstanding performance is likely a result of comprehensive data augmentation to tackle class imbalance, efficient transfer learning strategies, and the inherent strengths of the DenseNet121 architecture. These strengths include its deep layers and dense connections, which enable effective feature reuse.

TABLE III. Accuracy Comparison with Prior 7 Types of Skin Cancer Detection Studies

Ref.	Methodology	Accuracy (%)
[11], 2020	ResNet50	90
[12], 2022	MobileNetV2	93.11
[13], 2022	DenseNet201	82.9
[17], 2023	ViT	92.14
[20], 2023	EfficientNetB3	86
[15], 2024	InceptionV3 and DenseNet201	89
[16], 2024	ResNet101	92
[18], 2024	CNN	92.12
[19], 2024	ViT	93.75
[21], 2024	EfficientNet	85
This Study	**DenseNet121**	**94**

VI. Conclusion

This study demonstrates the effectiveness of transfer learning and Vision Transformers (ViTs) in improving the accuracy of skin cancer detection. By addressing the challenges of imbalanced data through extensive data augmentation, we achieved notable results with DenseNet121 and deepViT models, attaining accuracies of 94% and 92%, respectively. These findings underscore the potential of these advanced models in dermatological diagnostics. However, limitations persist, such as the need for large, diverse datasets to generalize across various skin tones and cancer types. Additionally, the computational demands of training ViTs and other deep-learning models remain significant. Future work will focus on overcoming these limitations by exploring hybrid approaches

that integrate the strengths of convolutional neural networks and transformers.

ACKNOWLEDGMENT

The authors acknowledge with thanks the technical support from the Faculty of Computing and Information Technology at the King Abdulaziz University (KAU), Jeddah, Saudi Arabia.

REFERENCES

[1] "Skin cancer," https://my.clevelandclinic.org/health/diseases/15818-skin-cancer, accessed: 20-08-2024.

[2] "Explore cancer statistics," American Cancer Society, https://cancerstatisticscenter.cancer.org/, 2024, accessed: 10-07-2024.

[3] "Cancer stat facts: Melanoma of the skin," National Cancer Institute, https://seer.cancer.gov/statfacts/html/melan.html, accessed: 05-07-2024.

[4] I. Ahmad, F. AlQurashi, E. Abozinadah, and R. Mehmood, "A novel deep learning-based online proctoring system using face recognition, eye blinking, and object detection techniques," *International Journal of Advanced Computer Science and Applications*, vol. 12, no. 10, 2021.

[5] I. Ahmad, F. Alqurashi, E. Abozinadah, and R. Mehmood, "Deep journalism and deepjournal v1. 0: a data-driven deep learning approach to discover parameters for transportation," *Sustainability*, vol. 14, no. 9, p. 5711, 2022.

[6] F. Alqurashi and I. Ahmad, "Scientometric analysis and knowledge mapping of cybersecurity." *International Journal of Advanced Computer Science & Applications*, vol. 15, no. 3, 2024.

[7] ——, "A data-driven multi-perspective approach to cybersecurity knowledge discovery through topic modelling," *Alexandria Engineering Journal*, vol. 107, pp. 374–389, 2024.

[8] I. Ahmad, F. AlQurashi, and R. Mehmood, "Machine and deep learning methods with manual and automatic labelling for news classification in bangla language," *arXiv preprint arXiv:2210.10903*, 2022.

[9] ——, "Potrika: Raw and balanced newspaper datasets in the bangla language with eight topics and five attributes," *arXiv preprint arXiv:2210.09389*, 2022.

[10] I. Ahmad and F. Alqurashi, "Early cancer detection using deep learning and medical imaging: A survey," *Critical Reviews in Oncology/Hematology*, p. 104528, 2024.

[11] H. K. Kondaveeti and P. Edupuganti, "Skin cancer classification using transfer learning," in *2020 IEEE International Conference on Advent Trends in Multidisciplinary Research and Innovation (ICATMRI)*, 2020, pp. 1–4.

[12] P. P. Naik, B. Annappa, and S. Dodia, "An efficient deep transfer learning approach for classification of skin cancer images," in *International Conference on Computer Vision and Image Processing*. Springer, 2022, pp. 524–537.

[13] M. Fraiwan and E. Faouri, "On the automatic detection and classification of skin cancer using deep transfer learning," *Sensors*, vol. 22, no. 13, p. 4963, 2022.

[14] N. Swetha R, V. K. Shrivastava, and K. Parvathi, "Multiclass skin lesion classification using image augmentation technique and transfer learning models," *International Journal of Intelligent Unmanned Systems*, vol. 12, no. 2, pp. 220–228, 2024.

[15] P. Vishnu, S. Krishvadana, C. Rani, R. Khoodeeram, A. Bouridane, and M. Rajesh Kumar, "Melanoma detection with transfer learning," in *2024 3rd International Conference on Artificial Intelligence For Internet of Things (AIIoT)*, 2024, pp. 1–6.

[16] K. K. Kaveti, M. S. Ravali, V. B. Sravya, and B. Deepthi, "Advancements of skin cancer classification using transfer learning segmentation," in *2024 3rd International Conference on Applied Artificial Intelligence and Computing (ICAAIC)*, 2024, pp. 612–616.

[17] M. A. Arshed, S. Mumtaz, M. Ibrahim, S. Ahmed, M. Tahir, and M. Shafi, "Multi-class skin cancer classification using vision transformer networks and convolutional neural network-based pre-trained models," *Information*, vol. 14, no. 7, p. 415, 2023.

[18] T. Tuncer, P. D. Barua, I. Tuncer, S. Dogan, and U. R. Acharya, "A lightweight deep convolutional neural network model for skin cancer image classification," *Applied Soft Computing*, p. 111794, 2024.

[19] G. Yang, S. Luo, and J. Li, "Advancing skin cancer classification across multiple scales with attention-weighted transformers," in *Fourth Symposium on Pattern Recognition and Applications (SPRA 2023)*, vol. 13162. SPIE, 2024, pp. 30–35.

[20] M. Ashfaq and A. Ahmad, "Skin cancer classification with convolutional deep neural networks and vision transformers using transfer learning," in *Advances in Deep Generative Models for Medical Artificial Intelligence*. Springer, 2023, pp. 151–176.

[21] N. Sánchez-Medel, V. Romero-Bautista, R. Díaz Hernández, and L. Altamirano Robles, "Comparison of cnns and vits for the detection of human skin lesions," in *Mexican Conference on Pattern Recognition*. Springer, 2024, pp. 274–283.

[22] P. Tschandl, "The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions," 2018. [Online]. Available: https://doi.org/10.7910/DVN/DBW86T

[23] N. Codella, V. Rotemberg, P. Tschandl, M. E. Celebi, S. Dusza, D. Gutman, B. Helba, A. Kalloo, K. Liopyris, M. Marchetti *et al.*, "Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic)," *arXiv preprint arXiv:1902.03368*, 2019.

[24] D. Zhou, B. Kang, X. Jin, L. Yang, X. Lian, Z. Jiang, Q. Hou, and J. Feng, "Deepvit: Towards deeper vision transformer," *arXiv preprint arXiv:2103.11886*, 2021.