An Improved Vision-Transformer Network for Skin Cancer Classification

Gayathri Mol Shajimon*, Isreal Ufumaka*, and Haider Raza*
* School of Computer Science and Electronics Engineering, University of Essex, Colchester, United Kingdom.
Email: gi22846@essex.ac.uk; iu22953@essex.ac.uk; h.raza@essex.ac.uk

***Abstract*—The early detection of skin cancer through automation is crucial for enhancing patient recovery prospects. In this study, we present an innovative approach for classifying skin cancer lesions using Vision transformers (ViTs) and evaluate it on the International Skin Imaging Collaboration (ISIC) 2017 dataset. The evolution of computer vision has led to the emergence of ViTs, which possess a unique ability to detect intricate patterns and features through self-attention mechanisms. This allow ViTs to recognize extensive dependencies within images, resulting in performance exceeding conventional CNN models. In comparison with the current state-of-the-art Inception-ResNet-V2 + Soft Attention (IRV2 + SA) technique, our proposed model exhibits superiority in accuracy, precision, recall, and AUC-ROC score for binary classification tasks in the ISIC 2017 challenge. Furthermore, the method demonstrates robustness and generalization capabilities, reinforcing its credibility as a reliable tool for lesion classification. The outcomes underscore ViTs' potential as a promising alternative to established convolutional neural network architectures for skin cancer lesion categorization.**

***Index Terms*—Vision Transformers, Skin Cancer Detection, Melanoma, Computer Vision, Focal Loss.**

I. INTRODUCTION

Skin cancer is a prevailing form of cancer and poses a significant public health concern worldwide due to the incidence, mortality rates and morbidity. Long-term exposure to ultraviolet radiation from the sun is considered to be the primary etiologic agent in the growth of skin cancer [1]. Compared to other cancers, skin cancer detection is challenging as it can sprout anywhere in the body. Among different types of skin cancer Melanoma is considered the most aggressive and deadliest[2]. The unique features of melanoma such as uneven distribution, asymmetrical shape, scalloped or notched borders, and uneven distribution of colours will be helpful in distinguishing it from other skin cancer types. Early identification of Melanoma is critical as it has the tendency to spread quickly to other parts of the body and is resistant to traditional treatments. Early detection of melanoma improves the chances of successful treatment and increases the possibility of patient survival. To handle this paramount need, the integration of Artificial Intelligence (AI) and machine learning has emerged as a reassuring approach to melanoma detection. Researchers have made significant progress in building automated and accurate methods to assist doctors in quickly and reliably diagnosing melanoma lesions by leveraging AI algorithms and computer vision techniques[3]. Recent studies show these new technologies outperform dermatologists in various multi-class skin cancer classifications [4].

The use of automatic skin cancer diagnosis has come a long way, progressing from traditional image processing methods to cutting-edge deep-learning models. Handcrafted features and rule-based algorithms were used in early efforts, but their inability to handle complicated skin lesions drove the introduction of data-driven methods[5]. CNN made a significant advance in capturing detailed features in skin lesions[6], with models such as VGGNet[7], ResNet [8], and InceptionNet[9] demonstrating outstanding performance[10]. CNN, on the other hand, relied on established spatial hierarchies, which gave rise to ViT. ViT excelled in capturing global and local dependencies by leveraging self-attention mechanisms, making it well-suited for analysing a variety of skin lesions [11]. ViT achieved state-of-the-art accuracy, sensitivity, specificity, and AUC when combined with large-scale skin lesion datasets such as International Skin Imaging Collaboration (ISIC), underlining its significance in furthering automated skin cancer identification. Ongoing research focuses on improving ViT structures and training methodologies to improve performance in skin cancer diagnosis.

In this paper, our aim is to explore the application of ViT in the field of skin cancer lesion detection, specifically in the ISIC 2017 dataset. The ISIC dataset offers a wide range and comprehensive collection of dermoscopic pictures, enabling researchers to create and test reliable skin cancer detection models[12]. We investigated the potential advantages and limitations of ViTs over current state-of-the-art Inception ResNetV2[10] comparing their performance. Our study intends to add to the growing body of knowledge in automated skin cancer detection by utilising ViT and the ISIC 2017 dataset. ViTs are based on the transformer architecture which was initially proposed for natural language processing tasks. In the field of machine translation, sentiment analysis and text generation, transformers have achieved groundbreaking success. This identical idea was applied to tasks in computer vision, leading to the emergence of ViTs. ViTs can directly analyze images, eliminating the requirement for pre-established spatial structures[13]. The findings from this study hold the promise to enhance the accuracy, efficiency, and accessibility of skin cancer diagnosis, ultimately benefiting patient outcomes and reducing healthcare system burdens.

This paper is organized as follows: Section II presents a background study, which highlights the recent advancements in deep learning in the field of skin cancer detection and the description of the dataset. Section III contains the methodology and proposed system which includes preprocessing techniques, the system architecture and solutions. Section IV contains the result and impact which includes all the experimental results and comparison with current state-of-the-art results. Sections V and VI contain discussions and conclusions regarding the experiment.

II. Backgroud

Recent times have witnessed significant strides in skin cancer detection research, encompassing traditional medicine, computer-aided image processing, deep learning, transfer learning, and Vision transformers (ViTs).

A. Traditional Methods and Technologies

Throughout history, traditional medicine has held a vital role in diagnosing diverse ailments, including skin cancer. Conventional techniques and technologies have been instrumental in disease diagnosis. According to Dorrell et al. (2019)[14], dermoscopy has been proposed as an economical method for dermatologists to rapidly assess suspicious lesions, enabling enhanced visual evaluation of lesion patterns and structures. Reported outcomes indicate that visual assessment without magnification attains a specificity of 0.81 and sensitivity of 0.71. With dermoscopy, medical experts can achieve specificity and sensitivity of 0.9 each. Despite dermoscopy's evolution, no increase in melanoma survival rates has been observed by Narayanamurthy in 2018[15]. Proficiency demands substantial training, and dermoscopy's subjective nature allows its combination with conventional methods like biopsy and histopathology examination. Additional skin cancer detection technologies, such as multiphoton tomography, high-frequency ultrasound, and pigmented skin lesion assay, have been catalogued by Dorrell et al[14]. Almaraz-Daminan et al. (2018)[15] highlighted variations in outcomes from diverse skin tests, advocating for an improved computer-aided approach. The suggested system should be compact, swiftly diagnostic, cost-effective, comfortable (biopsy-free), and boast high precision and sensitivity.

B. Computer-aided Image Processing

Computer-aided image processing is one promising field, as it allows for the analysis of a large amount of image data, which can be helpful in detecting melanoma cases quickly. Using a computer-aided approach, Jain et. al[16] applied the ABCD (i.e., Asymmetry, Border, Colour, Diameter) rule of dermoscopy, image preprocessing (gamma correction, resizing, and compensation of non-uniform illumination), segmentation, and feature extraction to give a lesion image analysis tool with the capability to detect melanoma. Saba T [17] compared between handcrafted features (i.e., texture-based Gabor wavelet transformation, local binary patterns, ABCD rule, seven-point checklist, Menzies scoring, etc) and non-handcrafted features (i.e., CNN, DCNN, transfer learning, and handcrafted features alongside local binary patterns combined with deep learning, etc.) popular for skin cancer diagnosis. It was posited that a well-trained computerized system can diagnose skin cancer without human intervention. Hair, glare, and shading removal from images, feature extraction for colours, texture and shape can be leveraged and then segmentation can be done using a combination of Watershed, Otsu and modified Otsu segmentation techniques [18]. Images are first properly attended to and high attention is given before these images are passed to a computer-aided classifier rather than a manual one.

C. Convolutional Neural Networks (CNN)

CNN has been a powerful tool for image recognition given its ability to analyse complexities in images. CNN has performed well when trained on a large enough dataset. Fu'adah et. al[19] proposed a fully-connected layered CNN model capable of multi-class classification (classes: dermatofibroma, nevus pigmentosus, melanoma, and squamous cell carcinoma) and tested it on the ISIC dataset. Similar to other image classification models, data augmentation was applied to balance the data resulting in 4,000 images in total used in training the predictive model. Results from the model showed the effectiveness of a modern computer-aided approach such as CNNs, as it attains 0.99 accuracy. Esteva et. al[6] also used a CNN model, showing its ability to classify skin cancer images. 129,450 clinical images were used and two cases are considered - keratinocyte carcinomas vs benign seborrheic keratoses and malignant melanoma vs benign nevi. In both cases, the results recorded are on par with a level of competence comparable to dermatologists. Subramanian et. al[20] were able to further demonstrate the potential of CNNs for skin cancer classification as their model was able to keep the false negative rate below 0.1, accuracy and precision above 0.8 when experimented on the HAM10000 dataset.

D. Transfer Learning

There is a need to use a high-performer model trained on another domain with easily obtained data due to data unavailability in a specific field [21]. Dorj et. al[22] considered four types of skin cancers (classes: actinic keratoses, basal cell carcinoma, squamous cell carcinoma, and Melanoma) from 3,753 images. This was applied to contemporary deep neural network techniques with ECOC SVM for classification coupled with feature extraction using AlexNet CNN. Using transfer learning, a small dataset was used (768 for train and 190 for test) to build a skin cancer detection model. Even with small data, an accuracy of 0.94, a sensitivity of 0.98, and a Specificity of 0.91 was achieved. Ali et. al[4] considered hair removal as a preprocessing technique on the HAM10000 dataset after which resizing and augmenting were done. The CNN EfficientNets B0-B7 was trained and evaluated on standard performance metrics such as. EfficientNet B4 and B5 gave the best results with ROC AUC scores of 0.98 and 0.98, although they had intermediate complexities. These results showed that more complexity in the model doesn't guarantee better performance.

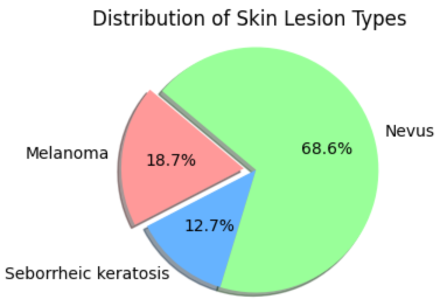

Fig. 1: Class distribution of ISIC 2027 training dataset

E. Vision image Transformers (ViTs)

Researchers have explored the potentials of ViTs for skin cancer classification given their ability to uncover long-range dependencies and complexities in images in ways CNNs are limited in. Yang et. al[23] proposes a novel ViT model for skin cancer lesion classification. Different techniques were used, such as rebalancing classes, splitting images into patches, tokenizing them, and using a transformer encoder with self-attention for classification. Using a pre-trained image net that was fine-tuned on the HAM10000 dataset, their results surpass that of Datta et. al[10] with an accuracy of 0.94. This shows that the attention layers enhanced the model's performance. Xin et. al[24] proposed a new ViT model for skin cancer image feature extraction and lesion classification using multi-scale patch embedding, overlapping sliding windows, and constructive learning. Applying this concept to the HAM10000 dataset, an accuracy of 0.94 and a precision of 0.94 was achieved.

F. ISIC 2017 DataSet

In this work, we have used ISIC 2017 dataset for the classification of skin cancer. The dataset has 2000 images in *.jpg* format and with image name in the format $ISIC\_ < image\_id > .jpg$, where image id is a unique 7-digit number. The ISIC 2017 challenge consists of 3 parts: 1) lesion segmentation; 2) lesion dermoscopic feature extraction; and 3) lesion classification. In this project, we focused mainly on lesion classification. In lesion classification, the aim is to perform binary classification, which involves 3 types of lesion classes, which are given as follows:

a) Melanoma - Malignant skin tumour (melanocytic)
b) Nevus - Benign cancer (melanocytic)
c) Seborrheic keratosis - A common noncancerous (benign) skin growth (non-melanocytic)

The distribution of the classes is illustrated in Fig. 1. Following are the two binary classification tasks:

1. Task 1: 'Melanoma' vs 'Nevus and Seborrheic Keratosis'
2. Task 2: 'Seborrheic Keratosis' vs 'Nevus and Melanoma'

III. Methodology and Proposed Solution

A. Problem Statement

In this work, we aim to explore novel approaches that can enhance the capability of automated skin cancer classification systems. ViT has the capability to capture global relationships and long-range dependencies within the images. Utilising ViTs, we aim to create a powerful and precise skin cancer classification system that outperforms current CNN-based state-of-the-art IRV2 + SA, thereby assisting dermatologists in the early and precise diagnosis of skin cancer lesions.

B. Methodology

1) Image Preprocessing: We resized our images to a size of 224×224 pixels to reduce the computational cost when the images are used in the model training. The Lanczos resampling filter method proposed by Vandame B [25] and Moraes et. al[26] is applied during the resizing process to improve image quality. The resampling method uses the Lanczos kernel function, denoted as $L(x)$, to perform the resampling. The Lanczos kernel defined in equation 1:

$$L(x) = \begin{cases} \text{sinc}(x) \cdot \text{sinc}\left(\frac{x}{a}\right) & \text{if } |x| \leq a \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

where a is a user-defined parameter known as the "filter size". The $\text{sinc}(x)$ function is given by $\frac{\sin(\pi x)}{\pi x}$. The Lanczos resampling process helps to preserve the visual integrity of the images and minimize artefacts that may arise due to resizing. It ensures that the images maintain their quality and provide accurate representations of the model during training and inference.

2) Data Augmentation: Usually, a large number of negative and positive samples are required for training the model to get better performance. As the dataset is small in size, we have applied a range data augmentation method to the images in the dataset to increase the number of images. Following are the data augmentation techniques we applied to the data set to increase the number of images in the dataset: *rotation range=180; width shift range=0.1; height shift range=0.1; zoom range=0.2; horizontal flip=True; vertical flip=True; and fill mode='nearest'*. Non-uniform transformation such as skewing and stretching has been avoided as it may cause non-uniform transformation. After augmentation, the number of images per class got increased to 5 times their original size.

3) Focal Loss: Focal Loss is generally used to handle the extreme imbalance issue between the two classes[27]. For the binary classification task, the focal loss starts from cross-entropy loss. The cross-entropy loss function is defined in the equation.2:

$$CE(p, y) = \begin{cases} -\log(p) & \text{if } y = 1, \\ -\log(1-p) & \text{otherwise.} \end{cases} \tag{2}$$

In the above $y \in \{\pm 1\}$ specifies the ground-truth class and $p \in [0, 1]$ is the model's estimated probability for the class with the label $y = 1$. For notational convenience, we define p_t as an equation. 3:

$$p_t = \begin{cases} p & \text{if } y = 1, \\ 1-p & \text{otherwise.} \end{cases} \tag{3}$$

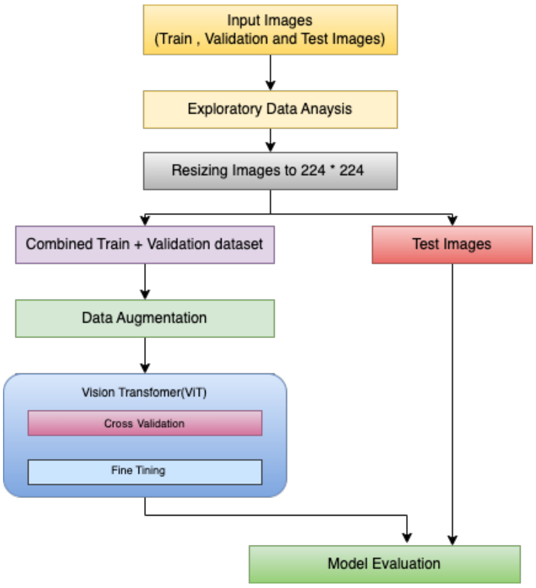

Fig. 2: Skin Cancer Detection System Block Diagram

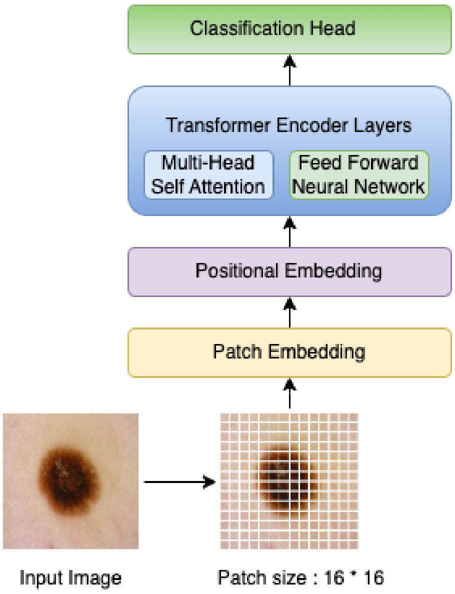

Fig. 3: General ViT architecture used for skin cancer classification.

We can rewrite the cross-entropy loss $CE(p, y)$ as $CE(p_t) = -\log(p_t)$. For handling the problem of imbalance, a new parameter α has been added in the above equation, which is termed as a balanced cross-entropy loss that is the base for focal loss. The equation 4 shown below defines the cross-entropy loss with p_t:

$$CE(p_t) = -\alpha_t \log(p_t). \tag{4}$$

For focal loss, [27] authors added new parameter γ to the above equation to handle extreme imbalance. In cross-entropy loss, negative samples dominate the gradient and contribute to the majority of the loss, leading to easier classification. α in the above equation for balanced cross-entropy balances the importance of positive and negative examples, it does not differentiate between easy or hard examples. As an alternative, they suggest reshaping the loss function to down-weight easy examples and concentrating training on difficult negatives by introducing γ in the above equation as follows 5::

$$FL(p_t) = -\alpha_t(1 - p_t)^{\gamma} \log(p_t) \tag{5}$$

where $-(1 - p_t)^{\gamma}$ is the modulating factor, in which γ is a tunable focusing parameter and $\gamma \geq 0$.

*4) **Proposed System**:* The Vision Transformer (ViT) model specifically the ViT-B16 variation is used in the proposed system for classifying skin cancer lesions. The skin cancer detection system block diagram is illustrated in Fig. 2. By incorporating self-attention processes, the ViT architecture pioneers a fresh method of image classification and equips the model to recognise intricate patterns and distant connections in the input images. The ViT architecture is illustrated in Fig. 3, which processes skin cancer input images by first dividing them into fixed-size patches (16*16). These patches are then linearly projected into high-dimensional vectors, capturing local image details. Positional embeddings are added to provide spatial information, and the patch embeddings are fed into a stack of transformer encoder layers. These layers use multi-head self-attention to capture global relationships and pass the embeddings through feed-forward neural networks for non-linearity. The final transformer encoder layer's output is globally averaged and passed through a fully connected layer with a softmax activation to obtain class probabilities for the input image. The foundation of the classification architecture in our implementation was the pre-trained ViT-B16 model. The pre-trained model was modified with additional layers to make it specifically suitable for the task of classifying skin cancer lesions. These layers included a flattened layer, layer normalization, drop out and dense layers. A 1-dimensional feature vector was created from the output of the ViT model using the flattened layer. The activations were normalised using layer normalisation, which also lessened the impact of internal covariate shift. The final classifier, the dense layers, assigned the collected features to the various subclasses of skin lesions (i.e., melanoma and non-melanoma).

The proposed architecture was trained using the Adam optimizer and optimized for focal loss. During training, the training data were fed to the model in batches, utilizing data generators for efficient memory utilization. Early stopping was employed to prevent overfitting, while a learning rate scheduler dynamically adjusted the learning rate throughout the training process. The suggested model illustrates its capability to accurately categorise skin cancer lesions through this architecture. Beyond the capabilities of traditional CNNs, the integration of the ViTs enables the capture of complex patterns and long-range dependencies. The proposed architecture is illustrated in Fig. 4. Data preprocessing for this experiment includes

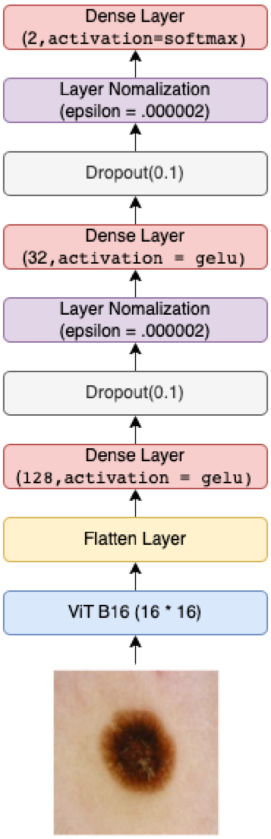

Fig. 4: Proposed 9-layer ViT Model Architecture with layer normalization and regularization

resizing images into fixed size, which is 224 * 224 in our case. Initially, we trained the model without resizing, but it took a minimum of 90 minutes to complete the training, but after resizing the computation time was reduced to 20 minutes. Data augmentation is the second preprocessing step we have implemented. In the data augmentation step, we excluded techniques such as stretching and skewing as this may cause a change in the shape of images which is an important feature classifying melanoma images, as the asymmetric border is a key feature in classifying melanoma images. We implemented hair removal and trained the image without hair. But there is no significant improvement in the result, so we trained without making many changes in the image quality.

*5) **Hardware and Software Requirements**:* The hardware used in the experiments includes an NVIDIA Tesla T4 GPU available on Google Colab with the following specifications: driver version 525.105.17 and CUDA version 12.0. The GPU has a memory capacity of 15,360MiB. We implemented our model using ViT-Keras 0.1.2 with TensorFlow version 2.12.0 as the backend, along with CUDA 12.0 for GPU acceleration. Our system is implemented with Python 3.10.6 on Google Colab.

*6) **Data Partitioning** :* The training dataset consists of 2000 images, where 3 classes are distributed as follows: Melanoma - 374 images; Seborrheic keratosis - 254 images; and Benign nevi - 1372 images. The class distribution is illustrated in Fig. 1. In addition to training data, separate validation (150 images) and test (600 images) are available along with their labels. In our study, we considered two different model evaluation settings: 1) S1: cross-validation (CV) and 2) S2: normal. Under each evaluation set, we have further divided it into two different data partitioning schemes: a) SS1: partitioned the training set into 5-folds and performed 5-fold CV and evaluated the model performance on the test set; b) SS2: merged the training and validation data then partitioned in into 5-folds and performed 5-fold CV and evaluated the model on the test set.

*7) **Model Training**:* Adam Optimizer was used for training the model with a learning rate of 0.00002 without exponential decay and we kept the default setting for hyperparameters β_1 and β_2, which control the exponential moving averages of the gradients and squared gradients, respectively. Model training was done for 20 epochs and prediction (i.e., model evaluation) was done on the unseen test set. Early stopping criteria have been configured with a patience of 4. For focal loss, we selected γ = 2 and α = .7 for handling the imbalance as recommended by [28].

*8) **Model Testing and Evaluation**:* For setting S1, 5-fold cross-validation (CV) was done for different combinations of binary classification with and without DA under two different data partitioning schemes (i.e., SS1 and SS2), given in Table. I and Table. II. For setting S2, normal training and test partitioning were done for different combinations of binary classification with and without DA under two different data partitioning schemes (i.e., SS1 and SS2), given in Table. III and Table. IV.

IV. Result

A. *Evaluation matrices*

In both setting S1 and S2, the test set (unseen data) was used to evaluate the model's performance. Due to the imbalance in the data, we have used a range of evaluation metrics such as recall, sensitivity, accuracy, and AUC-ROC score. Additionally, in the skin cancer detection model, detecting false negatives and achieving high recall scores is vital to ensure early identification of potential cases. This enhances the model's ability to capture a significant portion of actual positive cases, minimizing the risk of missing critical diagnoses. Furthermore, we also calculated, the AUC-ROC score as it quantifies the model's ability to distinguish between benign and malignant cases across various classification thresholds. A high AUC-ROC indicates effective discrimination, aiding in the selection of optimal models and ensuring accurate diagnostic outcomes.

The equations for accuracy, recall, and specificity are given below 6:

$$\begin{aligned} \text{accuracy} &= \frac{TP+TN}{TP+TN+FP+FN}, \\ \text{sensitivity/recall} &= \frac{TP}{TP+FN} \end{aligned} \tag{6}$$

TABLE I: Test dataset results for data partitioning scheme SS1 under setting S1 (i.e., training with 5-fold CV).

Model Name	Layer Normalization + FL				Batch Normalization + BCE				Batch Normalization + FL			
	Recall	Precision	Acc	AUC	Recall	Precision	Acc	AUC	Recall	Precision	Acc	AUC
Nev vs Seb (DA)	**0.90**	**0.91**	**0.90**	**0.94**	0.87	0.89	0.87	0.92	0.74	0.85	0.74	0.86
Nev vs Seb	**0.87**	0.87	**0.87**	**0.89**	0.75	0.85	0.74	0.86	**0.87**	**0.92**	0.75	0.81
Seb vs [Mel & Nev] (DA)	**0.89**	**0.89**	**0.89**	**0.89**	0.81	0.89	0.81	0.89	**0.89**	**0.89**	**0.89**	**0.89**
Seb vs [Mel & Nev]	**0.86**	**0.87**	**0.84**	0.85	0.68	0.86	0.68	0.85	0.708	0.86	0.708	0.84

TABLE II: Test dataset results for data partitioning scheme SS2 under setting S1 (i.e., training with 5-fold CV).

Model Name	Layer Normalization + FL				Batch Normalization + BCE				Batch Normalization + FL			
	Recall	Precision	Acc	AUC	Recall	Precision	Acc	AUC	Recall	Precision	Acc	AUC
Mel vs [Nev & Seb] (DA)	**0.83**	0.82	**0.83**	**0.83**	**0.83**	**0.83**	**0.83**	0.81	0.79	0.81	0.80	0.76
Mel vs [Nev & Seb]	**0.83**	**0.80**	**0.81**	**0.78**	0.76	0.78	0.76	0.74	0.55	0.73	0.55	0.61
Seb vs [Mel & Nev] (DA)	**0.87**	**0.88**	**0.87**	**0.88**	0.79	0.88	0.79	0.87	0.81	0.85	0.81	0.81
Seb vs [Mel & Nev]	**0.87**	**0.88**	**0.87**	**0.88**	0.80	0.83	0.80	0.77	0.80	0.83	0.80	0.77

TABLE III: Comparison with the state-of-the-art results on test dataset for data partitioning scheme SS1 under setting S2 (i.e., training on whole training data), where DA - Data Augmentation, FL - Focal Loss, BCE -Binary Cross Entropy)

Model Name	Nevus VS Seb				[Seb] vs [Melanoma & Nevus]			
	Sensitivity	Specificity	ACC	AUC	Sensitivity	Specificity	ACC	AUC
IRV2 (SA)	0.95	0.71	0.90	0.94	0.93	0.69	0.90	0.94
ViT B16 (BN)*	0.93	0.38	0.72	0.83	0.69	**0.83**	0.71	0.86
ViT B16 (LN)*	0.93	0.58	0.88	0.86	**0.93**	0.58	0.88	0.86
ViT B16 (BN + BCE + DA)*	0.95	0.68	0.89	0.68	0.94	0.76	**0.91**	0.91
ViT B16 (LN + BCE + DA)*	0.94	0.71	0.90	0.90	0.95	0.67	**0.91**	0.91
ViT B16 (BN+ FL + DA)*	**0.95**	0.68	0.89	0.93	0.92	0.70	0.89	0.89
ViT B16 (LN+ FL + DA)	0.93	**0.84**	**0.92**	**0.95**	**0.96**	0.59	**0.91**	0.93

TABLE IV: Comparison with the state-of-the-art results on test dataset for data partitioning scheme SS2 under setting S2 (i.e., training on whole training data), where DA - Data Augmentation, FL - Focal Loss, BCE -Binary Cross Entropy)

Model Name	[Melanoma] vs [Nevus & Seb]				[Seb] vs [Melanoma & Nevus]			
	Sensitivity	Specificity	ACC	AUC	Sensitivity	Specificity	ACC	AUC
ResNet50	0.63	0.89	0.84	0.86	0.87	0.84	0.84	0.95
RAN50 2	0.62	0.91	0.85	0.85	0.88	0.86	0.86	0.94
SEnet50 3	0.62	0.90	0.85	0.86	0.86	0.87	0.86	0.95
ARL-CNN	0.66	0.89	0.85	0.88	0.87	0.87	0.87	0.96
ViT B16 (BN)*	0.42	0.87	0.77	0.74	0.77	0.72	0.77	0.81
ViT B16 (LN)*	0.61	0.86	0.83	0.78	0.91	0.57	0.86	0.87
ViT B16 (BN + BCE + DA)*	0.56	0.88	0.82	0.75	0.88	0.8	0.87	0.89
ViT B16 (LN + BCE + DA)*	0.60	0.88	0.84	0.82	0.89	0.86	0.88	0.93
ViT B16 (BN + FL + DA)*	0.51	0.90	0.81	0.81	0.89	0.78	0.88	0.91
ViT B16 (LN + FL + DA)*	**0.77**	0.37	**0.86**	0.83	**0.96**	0.62	**0.91**	0.92

B. Experiment Results

We have conducted two sets of experiments S1 and S2, for both experiments, we used data partition schemes SS1 and SS2 as described in section 3.6. Setting S1 (i.e., 5-fold CV) results are presented in Table. I and Table. II. Setting S2 (i.e., normal training test partitioning) results are presented in Table. III and Table. IV.

In the case of setting S1 under data partitioning scheme SS1, the results are given in Table.I, ViT with layer normalization plus focal loss provided the best recall score in three pairwise binary classifications (i.e., Nev vs Seb with DA is 0.90, Nev vs Seb without DA is 0.87, and Seb vs [Mel & Nev] with DA is 0.89. In the case of setting S1 under data partitioning scheme SS2, the results are given in Table.II, a similar pattern was observed, where ViT with layer normalization plus focal loss provided the best recall score in all four pairwise binary classifications (i.e., Nev vs Seb with DA is 0.83, Nev vs Seb without DA is 0.83, Seb vs [Mel & Nev] with DA is 0.87, and Seb vs [Mel & Nev] is 0.87.

In the case of setting S2 under data partitioning scheme SS1, the results are given in Table.III, where simple training and test partitioning were done and compared the result of current state-of-the-art[10]. Evaluating nevus and Seborrheic Keratosis, we achieved a state-of-the-art recall score, with a performance improvement of 2% weighted recall and 1% AUC score, and also got higher specificity of 84% as well. In the case of Seb vs [Melanoma & Neves], we achieved a 0.96 recall score which is higher than the results obtained in [10].

In the case of setting S2 under data partitioning scheme SS2, the results are given in Table.IV, where simple training and test partitioning were done and compared the result with [29] ARL-CNN, SEnet, ResNet, RAN14, where they have used the original tasks in ISIC 2017 challenge. We achieved higher sensitivity and accuracy than ARL-CNN in both tasks (i.e., [Melanoma] vs [Nevus & Seb] and [Seb] vs [Melanoma & Nevus. In [Melanoma] vs [Nevus & Seb], we achieved a higher sensitivity of 77% and accuracy of 86%, which is higher than all the other state-of-the-art models presented in the Table. IV. In task 2 (i.e., Seb vs [Melanoma & Neves]), we achieved a sensitivity of 96% and an accuracy of 91 % which is higher than ARL-CNN [29] by 8% in terms of sensitivity and 4% in terms of accuracy.

V. Discussion

Different experiments have been conducted to study the potency of ViT in skin cancer classification. We explored and implemented multiple predictive models such as CNN, CNN with soft Attention[10], the basic architecture of ViT, ViT with batch normalization[24], and ViT with layer normalization with additional regularization. The combination of the ViT with layer normalization and regularisation methods produced the most significant results. Different preprocessing strategies, including hair removal[30], resizing, and data augmentation, were explored to choose the optimal approach. Surprisingly, hair removal didn't produce the expected performance gains. Resized images to 224x224 dimensions to speed up the training. Lanczos kernel function has been used to preserve the image quality during resizing. Data augmentation has been used to increase the number of images in each class but maintained data imbalance to get better performance as the loss function used is focal loss. A rigorous 5-fold CV has been done in all the underlying scenarios. ViTs with layer normalization and regularization showed a consistent performance in each fold while other model give good performance in one or two folds but was not consistent giving a poor average result, especially in scenarios without data augmentation.

Some of the challenges that affect the performance of the model include interclass similarity and intra-class dissimilarity. For example, Melanoma and benign nevi sometimes exhibit visual similarities which makes the differentiation challenging especially in training with limited classes[3].

The mortality rate among individuals with dark skin tones is comparatively high due to delayed diagnosis[31]. It is even difficult for dermatologists to distinguish the same. Therefore, it is challenging at the same time, it is important to implement classification techniques focused on the datasets, which contain images with darker skin tones. In the future, we suggest testing the effectiveness of ViT on larger datasets in order to fully utilise their powers. Larger data repositories offer the possibility of improved feature extraction and performance, especially given the limitations of the ISIC 2017 dataset. Additionally, taking into account the aforementioned difficulties could greatly benefit the field of skin cancer lesions classification.

VI. CONCLUSIONS

This paper introduces a novel 9-layer Vision Transformer (ViT) model for skin cancer lesion classification, built upon the foundational ViT architecture with 8 additional tailored layers. To combat dataset imbalance, we adopt focal loss as the loss function. The model harmoniously combines transformer-based vision architecture with dense layers and advanced regularization techniques, yielding accurate classifications. Through extensive research, our ViT model outperforms state-of-the-art[10], showcasing superior sensitivity, recall, and accuracy in binary classification tasks ISIC Challenge 2017. The model achieved an accuracy of 91% and 92% showing a performance improvement by 1% and 2% in task 1 and task 2 of the binary classification, respectively compared to [10]. Employing rigorous 5-fold cross-validation bolsters the consistency and robustness of our findings, affirming the model's reliability and potential to set new standards in accurate medical image classification.

ACKNOWLEDGMENT

HR was supported by the Economic and Social Research Council (ESRC) funded Business and Local Government Data Research Centre under Grant ES/S007156/1.

References

[1] D. L. Narayanan, R. N. Saladi, and J. L. Fox, "Ultraviolet radiation and skin cancer," *International journal of dermatology*, vol. 49, no. 9, pp. 978–986, 2010.

[2] P. T. Bradford, "Skin cancer in skin of color," *Dermatology nursing/Dermatology Nurses' Association*, vol. 21, no. 4, p. 170, 2009.

[3] M. Goyal, T. Knackstedt, S. Yan, and S. Hassanpour, "Artificial intelligence-based image classification methods for diagnosis of skin cancer: Challenges and opportunities," *Computers in biology and medicine*, vol. 127, p. 104065, 2020.

[4] K. Ali, Z. A. Shaikh, A. A. Khan, and A. A. Laghari, "Multiclass skin cancer classification using EfficientNets–a first step towards preventing skin cancer," *Neuroscience Informatics*, vol. 2, no. 4, p. 100034, 2022.

[5] J.-A. Almaraz-Damian, V. Ponomaryov, S. Sadovnychiy, and H. Castillejos-Fernandez, "Melanoma and nevus skin lesion classification using handcraft and deep learning feature fusion via mutual information measures," *Entropy*, vol. 22, no. 4, p. 484, 2020.

[6] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun, "Dermatologist-level classification of skin cancer with deep neural networks," *Nature*, vol. 542, no. 7639, pp. 115–118, 2017.

[7] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[8] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[9] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1–9.

[10] S. K. Datta, M. A. Shaikh, S. N. Srihari, and M. Gao, "Soft attention improves skin cancer classification performance," in *Interpretability of Machine Intelligence in Medical Image Computing, and Topological Data Analysis and Its Applications for Medical Data*. Springer, 2021, pp. 13–23.

[11] Y. Gulzar and S. A. Khan, "Skin lesion segmentation based on vision transformers and convolutional neural networks—a comparative study," *Applied Sciences*, vol. 12, no. 12, p. 5990, 2022.

[12] M. Berseth, "ISIC 2017-skin lesion analysis towards melanoma detection," *arXiv preprint arXiv:1703.00523*, 2017.

[13] K. Han, Y. Wang, H. Chen, X. Chen, J. Guo, Z. Liu, Y. Tang, A. Xiao, C. Xu, Y. Xu *et al.*, "A survey on vision transformer," *IEEE transactions on pattern analysis and machine intelligence*, vol. 45, no. 1, pp. 87–110, 2022.

[14] D. N. Dorrell and L. C. Strowd, "Skin cancer detection technology," *Dermatologic clinics*, vol. 37, no. 4, pp. 527–536, 2019.

[15] V. Narayanamurthy, P. Padmapriya, A. Noorasafrin, B. Pooja, K. Hema, K. Nithyakalyani, F. Samsuri *et al.*, "Skin cancer detection using non-invasive techniques," *RSC advances*, vol. 8, no. 49, pp. 28 095–28 130, 2018.

[16] S. Jain, N. Pise *et al.*, "Computer aided melanoma skin cancer detection using image processing," *Procedia Computer Science*, vol. 48, pp. 735–740, 2015.

[17] T. Saba, "Computer vision for microscopic skin cancer diagnosis using handcrafted and non-handcrafted features," *Microscopy Research and Technique*, vol. 84, no. 6, pp. 1272–1283, 2021.

[18] M. Vijayalakshmi, "Melanoma skin cancer detection using image processing and machine learning," *International Journal of Trend in Scientific Research and Development (IJTSRD)*, vol. 3, no. 4, pp. 780–784, 2019.

[19] Y. N. Fu'adah, N. C. Pratiwi, M. A. Pramudito, and N. Ibrahim, "Convolutional neural network (CNN) for automatic skin cancer classification system," in *IOP conference series: materials science and engineering*, vol. 982, no. 1. IOP Publishing, 2020, p. 012005.

[20] R. R. Subramanian, D. Achuth, P. S. Kumar, K. N. kumar Reddy, S. Amara, and A. S. Chowdary, "Skin cancer classification using convolutional neural networks," in *2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*. IEEE, 2021, pp. 13–19.

[21] K. Weiss, T. M. Khoshgoftaar, and D. Wang, "A survey of transfer learning," *Journal of Big data*, vol. 3, no. 1, pp. 1–40, 2016.

[22] U.-O. Dorj, K.-K. Lee, J.-Y. Choi, and M. Lee, "The skin cancer classification using deep convolutional neural network," *Multimedia Tools and Applications*, vol. 77, pp. 9909–9924, 2018.

[23] G. Yang, S. Luo, and P. Greer, "A novel vision transformer model for skin cancer classification," *Neural Processing Letters*, pp. 1–17, 2023.

[24] C. Xin, Z. Liu, K. Zhao, L. Miao, Y. Ma, X. Zhu, Q. Zhou, S. Wang, L. Li, F. Yang *et al.*, "An improved transformer network for skin cancer classification," *Computers in Biology and Medicine*, vol. 149, p. 105939, 2022.

[25] B. Vandame, "Fast and efficient resampling for multi-frame super-resolution," in *21st European Signal Processing Conference (EUSIPCO 2013)*. IEEE, 2013, pp. 1–5.

[26] T. Moraes, P. Amorim, J. V. Da Silva, and H. Pedrini, "Medical image interpolation based on 3D lanczos filtering," *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*, vol. 8, no. 3, pp. 294–300, 2020.

[27] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2980–2988.

[28] G. S. Tran, T. P. Nghiem, V. T. Nguyen, C. M. Luong, J.-C. Burie *et al.*, "Improving accuracy of lung nodule classification using deep learning with focal loss," *Journal of healthcare engineering*, vol. 2019, 2019.

[29] J. Zhang, Y. Xie, Y. Xia, and C. Shen, "Attention residual learning for skin lesion classification," *IEEE Transactions on medical imaging*, vol. 38, no. 9, pp. 2092–2103, 2019.

[30] A. Victor and M. R. Ghalib, "Automatic detection and classification of skin cancer." *International Journal of Intelligent Engineering & Systems*, vol. 10, no. 3, 2017.

[31] H. M. Gloster Jr and K. Neal, "Skin cancer in skin of color," *Journal of the American Academy of Dermatology*, vol. 55, no. 5, pp. 741–760, 2006.