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.

***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 a Vision transformer (ViT) and evaluate it on the International Skin Imaging Collaboration (ISIC) 2017 dataset. The evolution of computer vision has led to the emergence of ViT, which possesses a unique ability to detect intricate patterns and features through self-attention mechanisms. This allows ViT 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 t asks in the ISIC 2017 challenge. Furthermore, the method demonstrates 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. https://github.com/Gayathri-Shajimon/Skin-cancer-lesion-classification-using-ViT**

***Index Terms*—Vision Transformers, Melanoma, Focal Loss.**

I. INTRODUCTION

Skin cancer is a prevalent public health concern worldwide, which is primarily caused by chronic exposure to ultraviolet radiation from the sun [1]. Among them, melanoma is one of the deadliest and most aggressive types. Melanoma has unique features such as uneven distribution, asymmetrical shape, scalloped or notched borders, and uneven distribution of colours which helps us to differentiate from other types of skin cancers. Early detection of melanoma is crucial as it has the capability to spread rapidly and is resistant to traditional treatments. Machine learning combined with artificial intelligence and computer vision has demonstrated potential in melanoma detection, exceeding dermatologists across a range of classifications [2, 3]. Automatic skin cancer detection has evolved from traditional image processing to advanced deep-learning models. Hand-crafted features and rule-based algorithms were the earlier methods, but they couldn't handle complex skin lesions. Significant progress was made by CNNs [4], such as VGGNet, ResNet and InceptionNet in capturing detailed skin lesion features [5]. However, the introduction of ViT marked a breakthrough in capturing global and long-term dependencies using self-attention mechanisms, making them more suitable for various skin lesion detection [6]. ViTs have demonstrated remarkable performance when merged with datasets such as ISIC, underscoring their importance in automating the identification of skin cancer. The goal of our research is to improve training and ViT structures for more accurate skin cancer diagnosis.

In this paper, we aim to investigate applications of ViT in skin cancer lesion detection, focusing on the ISIC 2017 dataset, which provides a comprehensive collection of dermoscopic images for reliable model testing [7]. We compare ViTs with the current state-of-the-art Inception ResNetV2 [5] to compare their advantages and limitations. Our study contributes to the growing knowledge of automated skin cancer detection using ViT. The results hold the potential to enhance the accuracy, efficiency, and accessibility of skin cancer diagnosis, benefiting patient outcomes and reducing healthcare burdens.

The paper is organised as follows: Section II offers a background study on AI-assisted skin cancer detection. Section III outlines the methodology, including preprocessing, system architecture, and solutions. Section IV presents experimental results and comparisons. Section V discusses and concludes the experiments.

II. BACKGROUD

Almaraz-Daminan et al. (2018) [8] emphasised the need for non-invasive, low-cost computer-aided diagnostic instruments for the diagnosis of skin cancer. Dermatoscopy only slightly enhances the sensitivity and specificity of conventional procedures. However, the survival rate of melanoma is still quite low, and dermatoscopy requires a great deal of training. Using the ABCD rule by Jain et. al [9] and feature analysis by Saba T [10] in computer-aided image processing suggests the possibility of autonomous diagnosis. Esteva et. al [4] performed binary classification using a CNN model on 129,450 clinical images. Transfer Learning also exhibited great potential, as Dorj et. al [11] considered the multi-class classification of 3,753 skin images using deep neural network techniques with ECOC-SVM for classification coupled with feature extraction using AlexNet CNN. Using transfer learning on this small dataset they achieved an accuracy of 0.94, a sensitivity of 0.98, and a Specificity of 0.91. 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. Yang et. al [12] proposes a novel ViT model for skin cancer lesion classification. Using a pre-trained image net fine-tuned on the HAM10000 dataset, their results surpass that of Datta et. al [5] with an accuracy of 0.94. Xin et. al [13] 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 and precision of 0.94 was achieved.

III. Methodology and Proposed Solution

A. Methodology

1) ***Preprocessing****:* To manage computational costs during model training, images were resized to 224×224 pixels using the Lanczos resampling filter method. This approach maintained image quality, visual integrity, and minimized resizing artifacts. To mitigate the limited dataset size, data augmentation techniques were employed, including rotation (180°), width shift (0.1), height shift (0.1), zoom (0.2), horizontal and vertical flips, and fill mode ('nearest'). Non-uniform transformations such as skewing and stretching were deliberately avoided to prevent irregularities. Following augmentation, each class contained five times more images than the original dataset.

2) ***Focal Loss****:* Focal Loss is generally used to handle the extreme imbalance issue between the two classes [14]. The equation for focal loss is derived from the equation of cross-entropy loss by adding two parameters α and γ, where α which is termed as a balanced cross-entropy loss that is the base for focal loss, 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 below equation as follows 1

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

In the above equation, $p \in [0,1]$ is the model's estimated probability for the class.

3) ***Proposed System****:* The proposed skin cancer detection system utilizes the ViT model, specifically the ViT-B16 variant, which introduces self-attention mechanisms for recognizing complex patterns and long-range dependencies in input images. The ViT architecture processes skin cancer images by dividing them into fixed-size patches (16x16), projecting them into high-dimensional vectors, and adding positional embeddings for spatial information.These embeddings are then passed through transformer encoder layers with multi-head self-attention and non-linear feed-forward neural networks. The final output is globally averaged and processed through a fully connected layer with softmax activation for classification Fig. 1.

The architecture is built on the pre-trained ViT-B16 model and customized with flattened, layer normalization, dropout, and dense layers (see Fig. 2). The flattened layer generates a 1-dimensional feature vector, and layer normalization reduces the impact of internal covariate shift. Dense layers handle subclass classification. Data preprocessing, including image resizing to 224x224, significantly reduces computation time. Data augmentation retains crucial features like asymmetric

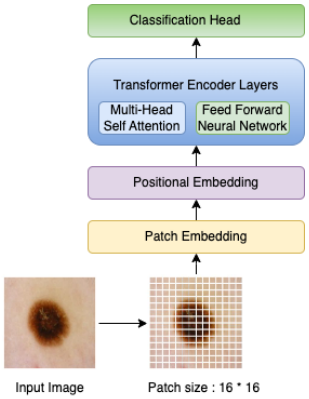

Fig. 1: ViT architecture used for skin cancer classification.

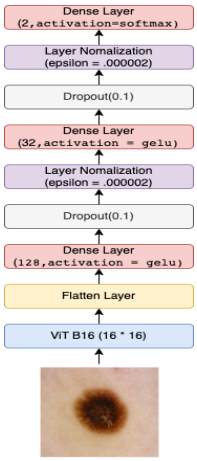

Fig. 2: Proposed 9-layer ViT Model Architecture with layer normalization and regularization

borders in melanoma images, deliberately excluding methods like stretching and skewing. The model is implemented using ViT-Keras 0.1.2, TensorFlow 2.12.0, and CUDA 12.0 for GPU acceleration in Python 3.10.6.

4) ***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. 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.

IV. Result

In both setting S1 and S2, the test set was used to evaluate the model's performance. Due to the imbalance in the data,

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	0.80	0.85	0.80	0.81

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

we have used a range of evaluation metrics such as recall, sensitivity, accuracy, and AUC-ROC score. In the context of setting S1 within data partitioning scheme SS1, the results are detailed in Table I. ViT with layer normalization plus focal loss stood out by providing the best recall scores in three pairwise binary classifications: Nevus vs. Seborrheic Keratosis with data augmentation (DA) achieved a recall of 0.90, Nevus vs. Seborrheic Keratosis without DA attained 0.87. Seborrheic Keratosis vs. [Melanoma & Nevus] with DA secured a recall of 0.89. Similarly, in setting S1 with data partitioning scheme SS2, the results can be found in Table II Here, we observed a similar pattern where ViT with layer normalization plus focal loss again outperformed other configurations, achieving the best recall score in all four pairwise binary classifications. This included a recall of 0.83 for Nevus vs. Seborrheic Keratosis with DA, a recall of 0.83 for Nevus vs. Seborrheic Keratosis without DA, a recall of 0.87 for Seborrheic Keratosis vs. [Melanoma & Nevus] with DA, and a recall of 0.87 for Seborrheic Keratosis vs. [Melanoma & Nevus] without DA. Shifting to setting S2 under data partitioning scheme SS1, the results are presented in Table III. In this configuration, simple training and test partitioning were employed, and a comparison with the current state-of-the-art [5] was conducted. Notably, we achieved a state-of-the-art recall score for Nevus and Seborrheic Keratosis, with a 2% improvement in weighted recall and a 1% increase in AUC score. Additionally, we obtained

a higher specificity of 84%. Under the same setting S2, with data partitioning scheme SS2, the results are detailed in Table IV. Here, we utilized simple training and test partitioning and compared the results with [15] ARL-CNN, SEnet, ResNet, and RAN14, which employed the original tasks from the ISIC 2017 challenge. Our model demonstrated superior sensitivity and accuracy compared to ARL-CNN in both tasks, outperforming all other state-of-the-art models. In the first task, [Melanoma] vs. [Nevus & Seborrheic Keratosis], we achieved a sensitivity of 77% and an accuracy of 86%, surpassing the other models. In the second task, [Seborrheic Keratosis] vs. [Melanoma & Nevus], we achieved a sensitivity of 96% and an accuracy of 91%, outperforming ARL-CNN by 8% in sensitivity and 4% in accuracy.

V. Discussion and Conclusion

Multiple experiments assessed the effectiveness of ViTs in skin cancer classification, comparing various models including CNN, CNN with soft attention, basic ViT, ViT with batch normalization, and ViT with layer normalization and regularization. ViT with layer normalization and regularization showed robustness. Different preprocessing techniques, such as data augmentation and resizing, were tested, addressing data imbalance while using focal loss as the loss function. A rigorous 5-fold CV was applied across all experiments. Challenges involved interclass similarity, intra-class dissimilarity, and difficulties distinguishing visually similar skin lesions, particularly with limited classes. Darker skin tones and delayed diagnosis presented additional challenges. Future research should incorporate larger datasets for enhanced performance and feature extraction. Tackling these issues is essential for advancing skin cancer lesion classification.

This paper presented an innovative 9-layer Vision Transformer (ViT) model for skin cancer lesion classification. It extended the foundational ViT architecture with 8 customized layers and addressed dataset imbalances using focal loss. The model synergized transformer-based vision architecture with dense layers and advanced regularization techniques, resulting in superior sensitivity, recall, and accuracy in binary classification tasks compared to the state-of-the-art [5]. Through rigorous 5-fold cross-validation, our ViT model exhibited reliability and potential for setting new standards in precise medical image classification, achieving a 1% and 2% performance improvement in task 1 and task 2 of binary classification, respectively, when compared to [5], with accuracy rates of 91% and 92%.

ACKNOWLEDGMENT

HR was supported by the Economic and Social Research Council (ESRC) funded Business and Local Government Data Research Centre (BLGDRC) under Grant ES/S007156/1. For the purpose of open access, the authors have applied a CC BY public copyright licence to any Author Accepted Manuscript (AAM) version arising from this submission.

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] 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.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

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

[8] 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.

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

[10] 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.

[11] 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.

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

[13] 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.

[14] 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.

[15] 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.