Enhancing Skin Lesion Classification: A Self-Attention Fusion Approach with Vision Transformer

Rahmat Izwan Heroza[0000−0001−7713−7556], John Q. Gan[0000−0003−1230−7643], and Haider Raza[0000−0002−3955−0144]

School of Computer Science and Electronic Engineering, University of Essex,
Colchester CO4 3SQ, United Kingdom
{rh22708,jqgan,h.raza}@essex.ac.uk

Abstract. Automated skin lesion classification is pivotal in modern dermatology, and recent strides in deep learning have shown immense potential in this field. This paper introduces a novel attention mechanism amalgamating various self-attention variants with the Vision Transformer (ViT) architecture to enhance skin lesion classification performance. By integrating Scaled Dot-Product Attention, Multiplicative Attention, and Additive Attention, a unified framework is devised for capturing diverse contextual cues within dermatological images. Extensive experiments are conducted on skin lesion datasets (ISIC 2017) to assess different loss functions, attention mechanisms, and fusion strategies. Results demonstrate that the proposed method significantly enhances classification performance across all metrics, exhibiting a remarkable improvement of over 12% in F1 score compared to the baseline. This approach not only showcases the efficacy of attention mechanisms in dermatological image analysis but also underscores the potential of ViT architecture in advancing automated skin lesion classification, thereby offering promising prospects for improving diagnostic accuracy and patient care in dermatology.

Keywords: Vision Transformer · Self-Attention Fusion · Skin Lesion Classification.

1 Introduction

Skin cancer analysis is a critical process used by medical professionals to detect and diagnose skin cancer [2]. This involves examining suspicious skin lesions and abnormalities to determine whether they are cancerous or benign. Various methods, including visual inspection, dermoscopy, and image analysis using artificial intelligence (AI), have been employed for accurate diagnosis. By identifying skin cancer at an early stage, medical interventions can be initiated promptly, leading to better chances of successful treatment and recovery [2]. Delays in detecting suspicious skin lesions significantly reduce five-year survival rates by 20%, as evidenced in a study [17]. The two-week rule for cancer patients requires immediate assessment of the suspected lesion by a specialist within two weeks.

Over the years, referrals for this pathway have surged, from 159,430 patients in 2009/2010 to 506,456 patients in 2019/2020, intensifying pressure on healthcare access and timely assessments [18]. Additionally, non-urgent referrals face an 18-week waiting period, with only 80% of patients seen within this timeframe in 2019/2020. The COVID-19 pandemic exacerbated this backlog by either cancelling appointments or accommodating urgent patients, further exacerbating access issues. With the ageing population, skin cancer referrals are projected to rise in the foreseeable future [12].

In the domain of skin cancer analysis, imbalanced datasets pose a significant challenge [13]. Such datasets often contain a disproportionate number of samples from certain types of skin lesions, resulting in an unequal representation of different skin cancer classes. This imbalance can adversely affect the performance of machine learning models and AI algorithms [13]. The skewed distribution may cause the model to be biased towards the majority class, leading to reduced accuracy in detecting and classifying rare or underrepresented skin cancer types. Addressing the imbalance problem is crucial to ensure the model's fairness and efficacy in skin cancer analysis. Vision Transformer (ViT) [11] is a state-of-the-art deep learning model that has shown remarkable performance in various computer vision tasks. However, ViTs demand immense labeled data for pre-training models, notably more than Convolutional Neural Networks (CNNs). Training ViTs from scratch on small datasets may yield suboptimal results. Transfer learning based on ImageNet presents challenges in adapting to domain-specific tasks with limited data. Furthermore, when applied to an imbalanced skin cancer image dataset, ViTs may encounter challenges due to the skewed class distribution [3]. The model may struggle to generalize effectively to the minority class, leading to suboptimal performance in detecting rare or malignant skin lesions. The need to enhance ViT's performance on imbalanced datasets is a pressing research topic to ensure accurate and unbiased skin cancer analysis using transformer-based models.

To overcome the limitations of ViTs on imbalanced skin cancer image datasets, this study proposes a method leveraging various techniques to mitigate class imbalance and enhance model performance: 1) employing data augmentation for synthetic minority class samples [22]; 2) utilizing weighted loss functions to emphasize rare classes during training [3]; and 3) modifying the ViT architecture by replacing the last n attention layers with attention fusion layers.

This paper is structured as follows: Section 2 describes the proposed network architecture with self-attention fusion and the loss functions. Section 3 provides information on the experiment setup, including dataset and evaluation metrics. Section 4 presents the results, followed by discussions. Section 5 presents future works, and Section 6 concludes the study.

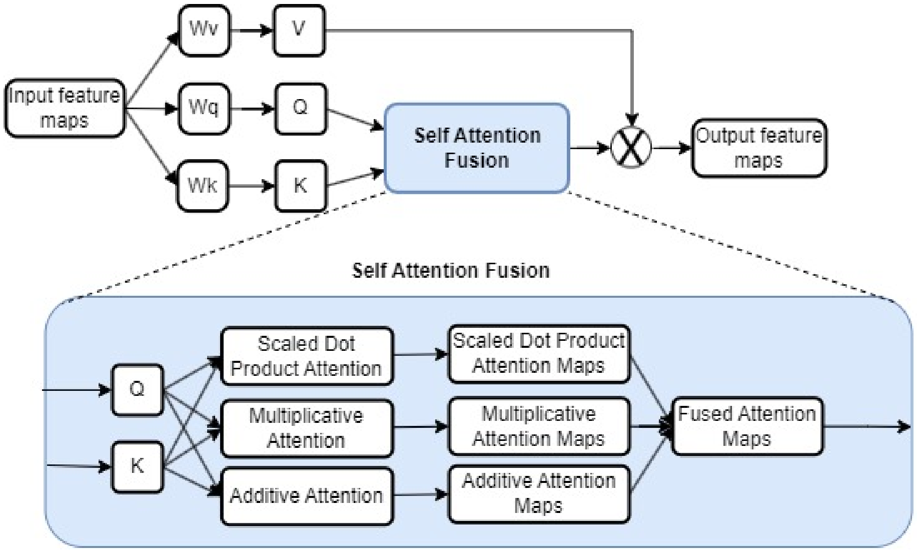


Fig. 1: The proposed self-attention fusion architecture

2 Method

2.1 Network Architecture with Self-attention Fusion

In our proposed method for skin cancer classification, ViT B16 [11] is adopted as the backbone of our model. This choice is underpinned by the superior performance and scalability demonstrated by ViTs in image recognition tasks. Two different sets of initial weights are used for the proposed model. The first one is the ViT model pre-trained via supervised learning on ImageNet-21k, a vast dataset with 14 million images across 21,843 classes, and subsequently fine-tuned on ImageNet-1k, comprising 1 million images spanning 1,000 classes. These weights are the same as those used in the original ViT paper [11]. The second one is the ViT model pre-trained via self-supervised learning on ImageNet-1k. This model follows the training method from the DINO paper [5]. This pre-training regimen equips our model with a rich understanding of visual features, enabling effective knowledge transfer to the domain of skin cancer classification. To enhance the model's generalization capabilities, RandAugment [8] is chosen for data augmentation.

To tailor the ViT model for dermatological image analysis, for capturing contextual information in particular, a modified attention layer is designed by combining the original scaled dot product attention [20] with other attention mechanisms such as multiplicative attention [16] and additive attention [4], as shown in Fig. 1. The scaled dot product attention, known for its efficiency and global context modeling, excels in capturing long-range dependencies by calculating the dot product of the query and key matrices while scaling to mitigate the vanishing gradient problem. Multiplicative attention, on the other hand, introduces a dynamic element-wise multiplication operation between the query and key matrices, emphasizing the interactive influence between features. Lastly, additive attention employs a weighted sum of the query and key matrices, allowing for a flexible and learnable combination of features.

The proposed attention fusion approach aims to leverage the strengths of different attention mechanisms and thus enhance the model's ability to discern salient features within dermatological images. Attention maps produced by different attention mechanisms are integrated using several fusion strategies namely average, maximum, and multiplication element-wise operations. The average fusion strategy as defined in (1) facilitates a balanced integration, allowing each attention mechanism to contribute equally to the final attention map. On the other hand, the maximum fusion strategy as defined in (2) emphasizes the most salient regions identified by any of the attention mechanisms, prioritizing the strongest cues. The multiplication fusion strategy as defined in (3), captures the synergistic effect of attention mechanisms, accentuating regions where multiple mechanisms concurrently identify salient features.

$$Avg\,Fusion(i,j) = \frac{1}{n}\sum_{k=1}^{n} \mathrm{Attmap}_k(i,j) \tag{1}$$

$$Max\,Fusion(i,j) = \max_{k} \mathrm{Attmap}_k(i,j) \tag{2}$$

$$Prod\,Fusion(i,j) = \prod_{k=1}^{n} \mathrm{Attmap}_k(i,j) \tag{3}$$

In our model, only the last n attention layers of the ViT model are strategically replaced with the proposed attention fusion layers, ensuring that the valuable knowledge encapsulated in the pre-trained weights from a vast dataset is retained. By focusing our adjustments on the last n attention layers, we aim to preserve the wealth of information encoded in the preceding layers while allowing the model to adapt and capture new intricate patterns specific to the task at hand. This fine-tuning methodology strikes a balance between leveraging the generalization capabilities gained from pre-training on a large dataset and tailoring the model to the nuances of the target domain, hence resolving the domain adaptation challenge. The modified attention layers serve as a lens, through which the proposed model can refine its understanding of the extracted features and effectively integrate domain-specific intricacies, thereby optimizing its performance on the specific tasks or patterns that may be characteristic of the new data distribution. In our experiments, we restrict the value of n to the set $\{1, 2, 3\}$ due to memory constraints.

2.2 Loss Functions

In order to combat the imbalance problem and optimize the performance of our skin cancer classification model based on the modified ViT architecture, we conducted extensive experiments to evaluate various loss functions, aiming to identify the most effective loss function for training the proposed model to achieve superior accuracy and generalization. The loss functions under scrutiny in our study include cross-entropy loss, as defined in (4), weighted cross-entropy loss, as defined in (5), with parameter weight $\alpha \in \{0.2, 0.5, 0.7, 0.8, 0.9\}$, focal loss [14], as defined in (6), with $\alpha \in \{0.2, 0.5, 0.7, 0.8, 0.9\}$ and parameter $\gamma \in \{1, 2\}$ which is a focusing parameter that modulates the loss. When $\gamma = 0$, Focal Loss becomes equivalent to standard cross-entropy loss. Additionally, we explored the class-balanced loss function [9], as defined in (7), with $\gamma \in \{0.5, 1\}$ and a hyperparameter that controls the balance between classes, $\beta \in \{0.999, 0.9999\}$. $\mathrm{p_t}$ in the equations represents the predicted probability assigned to the true label while $\mathrm{n_t}$ is the number of samples in the true class. Moreover, label smoothing [19] was incorporated into our experimentation with $\alpha \in \{0.1, 0.25, 0.5\}$.

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

$$Weighted\, CE(p_t) = -\alpha log(p_t) \tag{5}$$

$$FL(p_t) = -\alpha(1 - p_t)^{\gamma} log(p_t) \tag{6}$$

$$CB(p_t) = -\frac{1-\beta}{1-\beta^{n_t}}(1 - p_t)^{\gamma} log(p_t) \tag{7}$$

The weighted cross-entropy loss with varying alpha values allows us to assess the impact of different class weightings on the model's ability to prioritize certain classes during training. Similarly, the focal loss introduces a dynamic scaling factor, alpha, and a focusing parameter, gamma, to modulate the influence of hard-to-classify samples on the training process. The class-balanced loss, with adjustable beta and an additional gamma parameter, when applying the class-balanced term to the focal loss, addresses the issue of imbalanced class distribution, ensuring that each class contributes proportionally to the overall loss, hence preventing prior probability shift.

Label smoothing, on the other hand, is a regularization technique to prevent the model from becoming excessively confident in its predictions, potentially improving generalization performance. Through a meticulous analysis of the experimental results, we aim to provide valuable insights into the impact of these diverse loss functions on the ViT model's training dynamics and, ultimately, its efficacy in skin cancer classification.

3 Experimental Setup

3.1 Dataset and Performance Metrics

The International Skin Imaging Collaboration (ISIC) 2017 dataset is used in our experiment, which comprises a diverse collection of skin lesion images captured using dermoscopy photography. With ground truth annotations, it contains 2000 images in the training set, 150 images in the validation set, and 600 images in the test set. The images fall into one of three categories: "melanoma", "seborrheic keratosis", and "nevus" [7].

Due to the inherent imbalance in class distribution, the performance of our model is evaluated using a comprehensive set of metrics. The choice of metrics is essential to ensure a nuanced understanding of the model's effectiveness in handling the intricacies of imbalanced data. We employ the following metrics: Balanced Accuracy (BA), Area Under the Receiver Operating Characteristic curve (AUC), Average Precision (AP), Accuracy (Acc), and F1 Score (F1). These metrics collectively provide a holistic assessment of the model's classification performance, considering both its ability to correctly identify positive instances and its robustness in handling class imbalance.

BA is defined as:

$$BA = \frac{Sensitivity + Specificity}{2} \tag{8}$$

where Sensitivity, also known as True Positive Rate (TPR) or Recall, is calculated as follows:

$$Sensitivity = \frac{TP}{TP + FN} \tag{9}$$

and Specificity, also known as True Negative Rate (TNR), is calculated as follows:

$$Specificity = \frac{TN}{TN + FP} \tag{10}$$

It is a crucial metric that considers both sensitivity and specificity of the model, offering a fair assessment of the model performance on imbalanced datasets.

AUC quantifies the model's ability to discriminate between positive and negative instances across different probability thresholds. It is calculated as the area under the ROC curve.

AP measures the precision-recall trade-off across various threshold values, providing a more nuanced evaluation of classification performance, which is defined as:

$$AP = \sum_{n} (Recall_n - Recall_{n-1}) \times Precision_n \tag{11}$$

where $Recall_n$ and $Precision_n$ are Recall and Precision achieved at the n-th threshold.

Acc gauges the overall correctness of the model's predictions and is calculated as the ratio of correctly predicted instances to the total number of instances:

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} \tag{12}$$

F1 Score is the harmonic mean of precision and recall, offering a balanced assessment of the model's performance:

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{13}$$

3.2 Experimental Configuration

In our experimental study, distinct scenarios were systematically investigated, including variations in the choice of loss functions, as well as the impact of the replacement of the last attention layer with the proposed attention fusion. We adhere to the classification challenges outlined in the ISIC 2017 Challenge, encompassing two distinct tasks. The initial task (classification task 1) involves discriminating between (a) melanoma and (b) nevus and seborrheic keratosis, resulting in an imbalance ratio of 1:4. The subsequent task (classification task 2) requires distinguishing between (a) seborrheic keratosis and (b) nevus and melanoma, leading to an imbalance ratio of 1:7. To compare the performance of the proposed method, we use classification task 2 which has a larger imbalance ratio of 1:7. After that, we combine both classification tasks to compare the proposed method with those in the ISIC 2017 Challenge leaderboard. For each training instance, we employed ViT-B16 [11] as a baseline model. The baseline models consist of the original ViT architecture with pre-trained weights derived from both supervised and self-supervised learning, specifically DINO. Further, each model undergoes training on ISIC 2017 dataset for 20 epochs with a learning rate set to 2e-5 and AdamW [15] as the optimizer. The best model is determined based on the evaluation of the validation set in terms of the lowest evaluation loss. The reported performance metrics represent the average performance across the two tasks on the test set.

4 Results and Discussion

In the course of our investigation, we systematically scrutinized the performance of the proposed skin cancer classification model with different loss functions. The results of this initial comparative study are presented in Table 1, comparing the model performance on classifying (a) seborrheic keratosis and (b) a combination of nevus and melanoma.

Table 1: Performance Comparison of Different Loss Functions

Loss Function	**BA**	**AUC**	**AP**	**Acc**	**F1**
CE	0.6975	0.8535	0.5964	0.8467	0.5354
Weighted CE	0.6952	**0.8611**	**0.6103**	**0.8533**	**0.5368**
Focal Loss	0.5708	0.7868	0.4991	0.8200	0.2603
Class Balanced	**0.7166**	0.7794	0.4960	0.6950	0.4903

The weighted cross-entropy loss with $\alpha = 0.8$ achieved superior performance across all metrics except for BA. Building upon this preliminary result and considering that the ISIC 2017 Challenge utilizes the AUC as the main metric, the weighted cross-entropy loss was chosen to proceed with further investigation. The weighted cross-entropy loss is effective in handling class imbalance, which is often prevalent in medical imaging datasets such as the ISIC dataset.

In the subsequent study, we directed our attention toward enhancing the model's attention mechanism by modifying the last n attention layers. To deepen our understanding of the influence of attention mechanisms on the model's performance, we visualized the attention maps generated by different attention mechanisms, as depicted in Fig. 2b, where column 1 and column 3 show the attention maps on the penultimate layer from different heads, while column 2 and column 4 present the mask mapping of each attention map, obtained by selecting a portion of the self-attention maps through thresholding, retaining 30% of its mass. Visible parts are areas with attention values below the threshold, while the red areas indicate areas with attention values above the threshold that the model pays more attention to. It shows that each attention mechanism can identify both brighter and darker skin lesions in column 2 and column 4, respectively, albeit with varying intensity levels in column 1 and column 3. This divergence in intensity provides an opportunity to effectively amalgamate their respective attention maps.

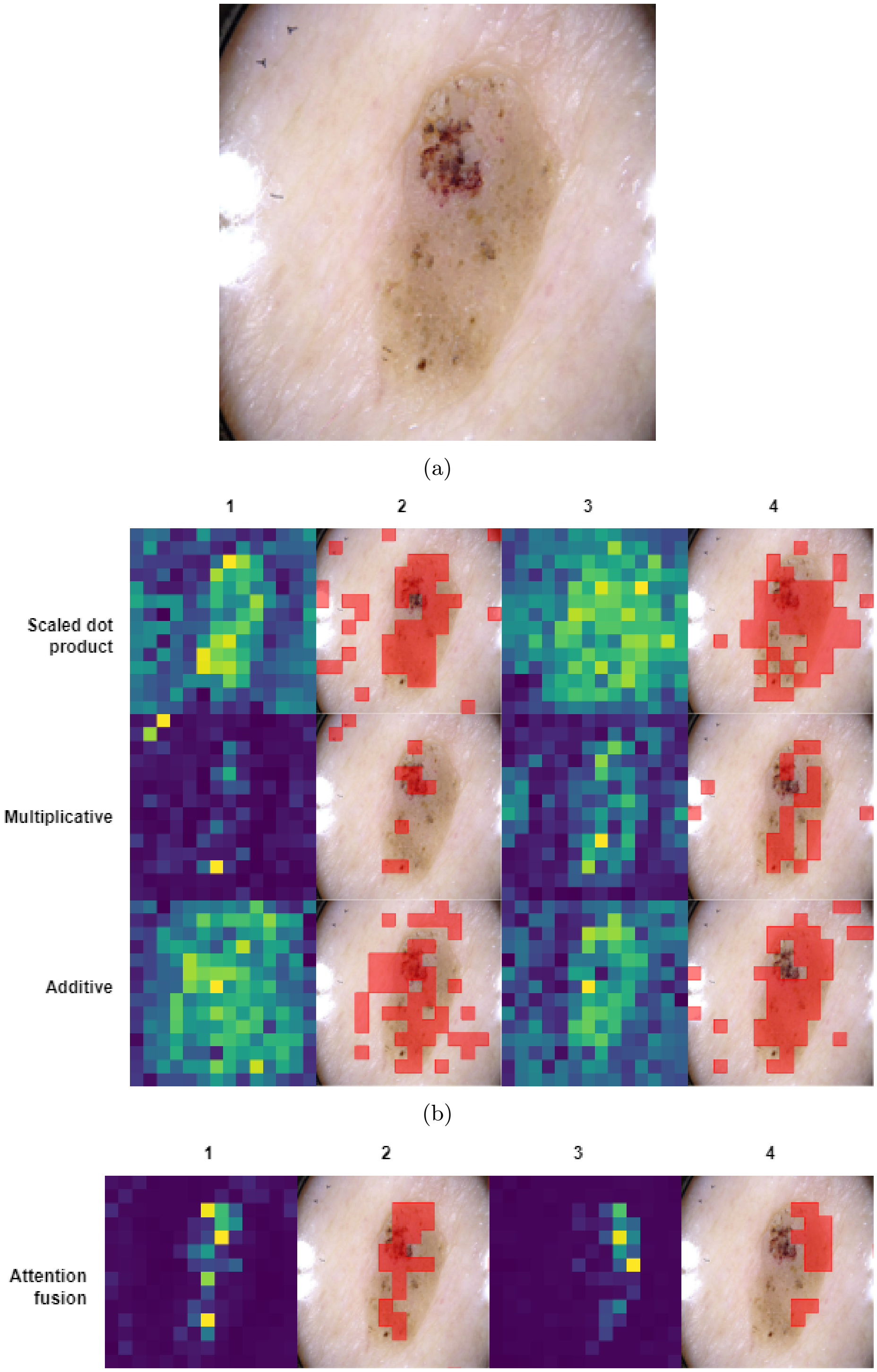
Subsequently, we explored the synergy achievable through attention map fusion in our proposed method. By combining attention maps produced by different attention mechanisms, we sought to create a more comprehensive and informative representation of salient features within dermatological images. We used the ISIC 2017 dataset to train our models, employing weighted cross entropy with a loss function, as recommended in the preceding phase.

The results of this attention mechanism modification are systematically presented in Table 2 for the ViT model and DINO model. It is evident that all the proposed models using the ViT model with different attention fusion strategies outperformed the baseline model in terms of the AP metric. The proposed model using the ViT model incorporating maximum attention fusion particularly outperformed all other models across all metrics, except in the case of AUC, in terms of which the product fusion approach outperformed all other methods.

Table 2: Self-attention Fusion Performance on VIT-B16 and DINO-B16

Fusion Strategy	**BA**	**AUC**	**AP**	**Acc**	**F1**
VIT-B16					
Base	0.6952	0.8611	0.6103	0.8533	0.5368
Prod. Fusion (Ours)	0.6899	**0.8658**	0.6277	0.8500	0.5263
Max Fusion (Ours)	**0.7005**	0.8465	**0.6305**	**0.8567**	**0.5474**
Avg Fusion (Ours)	0.6919	0.8449	0.6167	0.8533	0.5319
DINO-B16					
Base	0.6962	0.8543	0.6180	0.8550	0.5397
Prod. Fusion (Ours)	0.7698	0.8607	0.6448	0.8483	0.6224
Max Fusion (Ours)	**0.7768**	**0.8787**	**0.6516**	**0.8700**	**0.6518**
Avg Fusion (Ours)	0.7665	0.8714	0.6439	0.8483	0.6192

It is evident from Table 2 that our proposed attention fusion method is not only effective on the original ViT model, but also on the models pre-trained



(a)

(b)

(c)

Fig. 2: (a) Original image. (b) Attention maps from two different heads correspond to a skin lesion with a darker and brighter area. The even columns present the mask mapping with a threshold of 0.3. (c) Attention maps from the proposed method

using the DINO method. When applied on the models pre-trained using the DINO method, the proposed method incorporating maximum attention fusion shows the best performance in terms of all metrics, especially in terms of F1 score, with a 12% increase compared to the base model.

The initial assessment reveals that the base models, utilizing both VIT and DINO architectures, yield similar levels of performance. However, our proposed approach exhibits a significant performance boost when applied to the DINO model, revealing substantial disparities in performance. This outcome suggests that the intricate process of feature extraction, facilitated by the three attention mechanisms, operates with greater efficacy when applied to the features generated by the DINO architecture. These results align closely with the findings presented in the DINO paper [5], where it is elucidated that DINO's features harbor richer and more informative content essential for tasks such as semantic segmentation. The inherent self-distillation mechanism within the DINO framework which produced the final model plays a pivotal role in enhancing the model's ability to capture nuanced and meaningful features inherent within the skin lesion images.

The improvement in performance observed in the proposed model compared to the baseline could be attributed to its enhanced focus on the lesion area, as evidenced by the attention maps visualization in Fig. 2c. These attention maps reveal that the proposed model exhibits a greater degree of attention and emphasis on the features within the lesion area while tending to disregard irrelevant regions of the skin. This heightened focus on the lesion area likely allows the model to extract more discriminative features associated with the pathology, thereby improving its ability to accurately classify or detect abnormalities.

By selectively attending to the lesion area, the proposed model may effectively filter out the noise and irrelevant information present in the surrounding skin regions. This targeted attention mechanism enables the model to prioritize relevant features crucial for classification or detection tasks, leading to a more robust and accurate performance compared to the baseline. Additionally, the ability of the proposed model to ignore non-essential areas of the skin suggests a more efficient allocation of computational resources, allowing for a more refined analysis of the critical regions.

Furthermore, the attention maps provide valuable insights into the inner workings of the proposed model, shedding light on its decision-making process and highlighting areas of focus that contribute most significantly to its improved performance. Overall, the relationship between the observed performance improvements and the attention maps visualization underscores the importance of attention mechanisms in enhancing the effectiveness of deep learning models for medical image analysis tasks.

In order to measure the confidence of the classification performance of our proposed methods, the McNemar test was used, which is a statistical method commonly employed in the field of machine learning to assess the differences in classification performance between two models with low type I error [10]. It provides a robust means of evaluating whether the observed discrepancies in predictions made by the models are statistically significant. By examining the discordant classifications made by the models on paired data points, the McNemar test enables to determine if one model significantly outperforms the other.

The contingency tables and p values for comparing the proposed models with self-attention fusion and the baseline models VIT-B16 and DINO-B16 using McNemar Test are presented in Table 3 and Table 4 respectively. The null hypothesis in this test is that the proposed model with self-attention fusion and the baseline show error rates without significant difference. While our proposed method demonstrates some improvement over the baseline model when applied to the VIT architecture, the difference is not statistically significant. The test results fail to reject the null hypothesis, indicating similar error rates between the two methods. However, on the DINO architecture, our proposed method exhibits clear superiority. With a p-value of 0.0131, falling below the conventional threshold of 0.05, the test successfully rejects the null hypothesis, underscoring a significant enhancement in model performance when our proposed attention fusion method is employed with the DINO model.

Table 3: McNemar Test on VIT-B16

Contingency Table	Ours - Correct	Ours - Incorrect
Base - Correct	497	15
Base - Incorrect	17	71
p-value: 0.8596	**Fail to Reject Null Hypothesis**	

Table 4: McNemar Test on DINO-B16

Contingency Table	Ours - Correct	Ours - Incorrect
Base - Correct	476	22
Base - Incorrect	43	59
p-value: 0.0131	**Reject Null Hypothesis**	

In Table 5, our results were compared with those in the ISIC 2017 Challenge leaderboard [1] in terms of the average performance for classification task 1 and classification task 2.

In contrast to the outcomes observed in the ISIC 2017 Challenge leaderboard, our proposed method secures the 4th position based on the AUC metric, which serves as the primary evaluation criterion in the challenge. It is noteworthy that this accomplishment was attained despite our utilisation of ViT B16 as the base model, a choice made to accommodate memory constraints in our experimental

Table 5: Comparison with the ISIC 2017 Leaderboard

Model	**AUC**	**BA**	**AP**	**Acc**	**F1**
Rank 1	**0.911**	0.831	0.750	0.816	0.612
Rank 2	0.910	**0.883**	0.748	0.849	0.242
Rank 3	0.908	0.844	**0.754**	0.883	0.564
DINO Fusion (Ours)	0.899	0.805	0.695	**0.890**	**0.679**
Rank 4	0.896	0.843	0.733	0.888	0.612
VIT Fusion (Ours)	0.896	0.770	0.699	0.871	0.606
Rank 5	0.886	0.847	0.667	0.873	0.608

setup. It should be noted that the top three models in the leaderboard leverage more extensive training data and segmentation masks. It is important to highlight that the potential for achieving even more superior results exists by employing larger ViT variants [11] and incorporating a more extensive set of training data. It is worth noting that our proposed method outperformed all the models in the leaderboard in terms of Acc and F1 score, demonstrating an excellent balance between precision and recall in our model's performance.

5 Future Work

There exist different feature map dimensions, such as Linformer [21] and Performer [6]. It is worth exploring the synergistic potential of integrating the latest advancements in self-attention mechanisms in future investigations. Linformer and Performer have emerged as promising alternatives to traditional self-attention mechanisms, offering more efficient computation and scalability to handle larger input dimensions. By combining these novel attention maps with varying feature map dimensions, we aim to further enhance the performance and robustness of skin lesion classification models. This endeavor could entail exploring the interplay between different attention mechanisms and feature map structures, optimizing their fusion strategies, and conducting comprehensive evaluations on diverse dermatological datasets. Additionally, investigating the interpretability and generalisability of such hybrid models will be crucial for advancing our understanding of skin lesion classification and facilitating their clinical applicability in real-world settings.

6 Conclusion

This paper proposes a simple yet powerful attention fusion technique within a single-model framework, designed to enhance the skin cancer classification capabilities of vision transformers. By selectively replacing attention mechanisms in the final layers with a combination of diverse attention variants, our approach leverages pre-trained weights while exploring new patterns. Experimental results have demonstrated the effectiveness of the proposed method, with consistent improvements across all metrics, achieving a 12% increase in F1 score compared to the baseline. Notably, even with a small model due to memory constraints our proposed method attained the fourth position compared to those in the ISIC 2017 Challenge leaderboard in terms of AUC and outperformed all the models in the leaderboard in terms of Acc and F1 score.

References

1. Isic 2017 leaderboard, https://challenge.isic-archive.com/leaderboards/2017/
2. Anderson, A.M., Matsumoto, M., Saul, M.I., Secrest, A.M., Ferris, L.K.: Accuracy of skin cancer diagnosis by physician assistants compared with dermatologists in a large health care system. JAMA Dermatology **154**, 569–573 (5 2018). https://doi.org/10.1001/JAMADERMATOL.2018.0212
3. Ayas, S.: Multiclass skin lesion classification in dermoscopic images using swin transformer model. Neural Computing and Applications **35**, 6713–6722 (3 2023). https://doi.org/10.1007/S00521-022-08053-Z
4. Bahdanau, D., Cho, K.H., Bengio, Y.: Neural machine translation by jointly learning to align and translate. 3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings (9 2014). https://doi.org/10.48550/arXiv.1409.0473
5. Caron, M., Touvron, H., Misra, I., Jegou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. Proceedings of the IEEE International Conference on Computer Vision pp. 9630–9640 (4 2021). https://doi.org/10.1109/ICCV48922.2021.00951
6. Choromanski, K., Likhosherstov, V., Dohan, D., Song, X., Gane, A., Sarlos, T., Hawkins, P., Davis, J., Mohiuddin, A., Kaiser, L., Belanger, D., Colwell, L., Weller, A.: Rethinking attention with performers. ICLR 2021 - 9th International Conference on Learning Representations (9 2020). https://doi.org/10.48550/arXiv.2009.14794
7. Codella, N.C., Gutman, D., Celebi, M.E., Helba, B., Marchetti, M.A., Dusza, S.W., Kalloo, A., Liopyris, K., Mishra, N., Kittler, H., Halpern, A.: Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic). Proceedings - International Symposium on Biomedical Imaging **2018-April**, 168–172 (5 2018). https://doi.org/10.1109/ISBI.2018.8363547
8. Cubuk, E.D., Zoph, B., Shlens, J., Le, Q.V.: Randaugment: Practical automated data augmentation with a reduced search space. IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops **2020-June**, 3008–3017 (9 2019). https://doi.org/10.1109/CVPRW50498.2020.00359
9. Cui, Y., Jia, M., Lin, T.Y., Song, Y., Belongie, S.: Class-balanced loss based on effective number of samples. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition **2019-June**, 9260–9269 (1 2019). https://doi.org/10.1109/CVPR.2019.00949
10. Dietterich, T.G.: Approximate statistical tests for comparing supervised classification learning algorithms. Neural Computation **10**, 1895–1923 (10 1998). https://doi.org/10.1162/089976698300017197
11. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby,

N.: An image is worth 16x16 words: Transformers for image recognition at scale. ICLR 2021 - 9th International Conference on Learning Representations (10 2020). https://doi.org/10.48550/arXiv.2010.11929

12. Garbe, C., Amaral, T., Peris, K., Hauschild, A., Arenberger, P., Basset-Seguin, N., Bastholt, L., Bataille, V., Del Marmol, V., Dréno, B., et al.: European consensus-based interdisciplinary guideline for melanoma. part 1: Diagnostics: Update 2022. European Journal of Cancer **170**, 236–255 (2022). https://doi.org/10.1016/j.ejca.2022.03.008
13. Gessert, N., Sentker, T., Madesta, F., Schmitz, R., Kniep, H., Baltruschat, I., Werner, R., Schlaefer, A.: Skin lesion classification using cnns with patch-based attention and diagnosis-guided loss weighting. IEEE Transactions on Biomedical Engineering **67**, 495–503 (2 2020). https://doi.org/10.1109/TBME.2019.2915839
14. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollar, P.: Focal loss for dense object detection. IEEE Transactions on Pattern Analysis and Machine Intelligence **42**, 318–327 (8 2017). https://doi.org/10.1109/TPAMI.2018.2858826
15. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. 7th International Conference on Learning Representations, ICLR 2019 (11 2017), https://arxiv.org/abs/1711.05101v3
16. Luong, M.T., Pham, H., Manning, C.D.: Effective approaches to attention-based neural machine translation. Conference Proceedings - EMNLP 2015: Conference on Empirical Methods in Natural Language Processing pp. 1412–1421 (8 2015). https://doi.org/10.18653/v1/d15-1166
17. Pacifico, M., Pearl, R., Grover, R.: The uk government two-week rule and its impact on melanoma prognosis: an evidence-based study. The Annals of The Royal College of Surgeons of England **89**(6), 609–615 (2007). https://doi.org/10.1308/003588407x205459
18. Smith, L., Sansom, N., Hemphill, S., Bradley, S.H., Shinkins, B., Wheatstone, P., Hamilton, W., Neal, R.D.: Trends and variation in urgent referrals for suspected cancer 2009/2010–2019/2020. British Journal of General Practice **72**(714), 34–37 (2022). https://doi.org/10.3399/bjgp22X718217
19. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J.: Rethinking the inception architecture for computer vision. Proceedings of the IEEE conference on computer vision and pattern recognition pp. 2818–2826 (2016). https://doi.org/10.48550/arXiv.1512.00567
20. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Łukasz Kaiser, Polosukhin, I.: Attention is all you need. Advances in Neural Information Processing Systems **2017-December**, 5999–6009 (6 2017). https://doi.org/10.48550/arXiv.1706.03762
21. Wang, S., Li, B.Z., Khabsa, M., Fang, H., Ma, H.: Linformer: Self-attention with linear complexity (6 2020). https://doi.org/https://doi.org/10.48550/arXiv.2006.04768
22. Zhao, C., Shuai, R., Ma, L., Liu, W., Hu, D., Wu, M.: Dermoscopy image classification based on stylegan and densenet201. IEEE Access **9**, 8659–8679 (2021). https://doi.org/10.1109/ACCESS.2021.3049600