Integrating Anatomical Site Information into Vision Transformer Models for Skin Cancer Classification

Dongwon Woo[1], Jungrae Cho[1], Sumok Bae[3], Dae-Lyong HA, M.D., Ph.D.[2], Weon Ju Lee, M.D., Ph.D. [2], and Sungmoon Jeong[1,3*]

[1]Research Center for Artificial Intelligence in Medicine, Kyungpook National University Hospital, Daegu, South Korea

[2]Department of dermatology, School of medicine, Kyungpook National University

[3]Department of Medical Informatics, School of Medicine, Kyungpook National University, Daegu, South Korea

woodongwon23@gmail.com, zzemb6@gmail.com, jeongsm00@gmail.com

Abstract.

Early and accurate diagnosis of skin cancer is critical for improving patient outcomes. In this study, we propose a novel approach for classifying skin lesions into four categories: Benign, Basal Cell Carcinoma (BCC), Squamous Cell Carcinoma (SCC), and Melanoma, using a Vision Transformer (ViT)-based model. We further investigate the impact of incorporating site-specific information on the model's performance. The approach incorporates various methods to integrate anatomical site information with the visual features extracted by the ViT backbone. The combined data is then fed into a classifier to generate the final prediction. Experimental results demonstrate that including site information leads to a modest but consistent improvement in classification accuracy. This suggests that anatomical context provides valuable auxiliary information that enhances the model's ability to distinguish between different types of skin lesions. Our findings highlight the potential of integrating clinical metadata with advanced deep learning models to improve the diagnostic accuracy of skin cancer classification systems.

Ⅰ. Introduction

Skin cancer is the most common type of cancer globally, with rising incidence rates annually. Early and accurate diagnosis is essential for improving patient outcomes and reducing mortality [1,2]. Among skin cancers, Basal Cell Carcinoma (BCC), Squamous Cell Carcinoma (SCC), and Melanoma are the most prevalent and present significant challenges in differentiation due to their visual similarities, especially in early stages.

While convolutional neural networks (CNNs) have shown promise in automated skin lesion classification, they often struggle to capture global contextual information, limiting their generalizability. Vision Transformers (ViTs) have recently emerged as a powerful alternative, demonstrating superior performance in various computer vision tasks by effectively capturing long-range dependencies within images [3]. This capability makes them particularly well-suited for complex medical image analysis, including skin cancer classification.

However, the potential of integrating clinical metadata, such as the anatomical site of the lesion, remains underexplored. Anatomical site information is crucial, as different types of skin cancer tend to occur in specific body regions. Incorporating this context could enhance the model's accuracy, especially in cases where visual differences are minimal.

In this study, we propose a ViT-based model to classify skin lesions into four categories: Benign, BCC, SCC, and Melanoma. Additionally, we explore various methods for integrating anatomical site information with visual features extracted by the ViT backbone, hypothesizing that this integration will lead to improved classification performance.

Fig. 1 Examples of skin cancer

Ⅱ. Method

2.1 Dataset

We collected a total of 10,068 skin cancer images using three devices (smartphone, dermatoscope, and DSLR) from 2004 to 2024. For model building and performance testing, we defined two test sets. The first is assigned as ClinicalTrial set, which consists of 791 images selected by clinicians to meet the criteria for clinical trials. The second is named Demonstration set, composed of 542 images collected after June 2, 2023. The model training and validation sets were created by randomly splitting the remaining data (excluding each test set) into an 80:20 ratio. Among the entire dataset, there are 4,286 images of Benign cases, 2,580 images of BCC, 1,729 images of SCC, and 1,473 images of Melanoma. The anatomical sites are categorized into four types: acral, extremity, trunk, and head & neck.

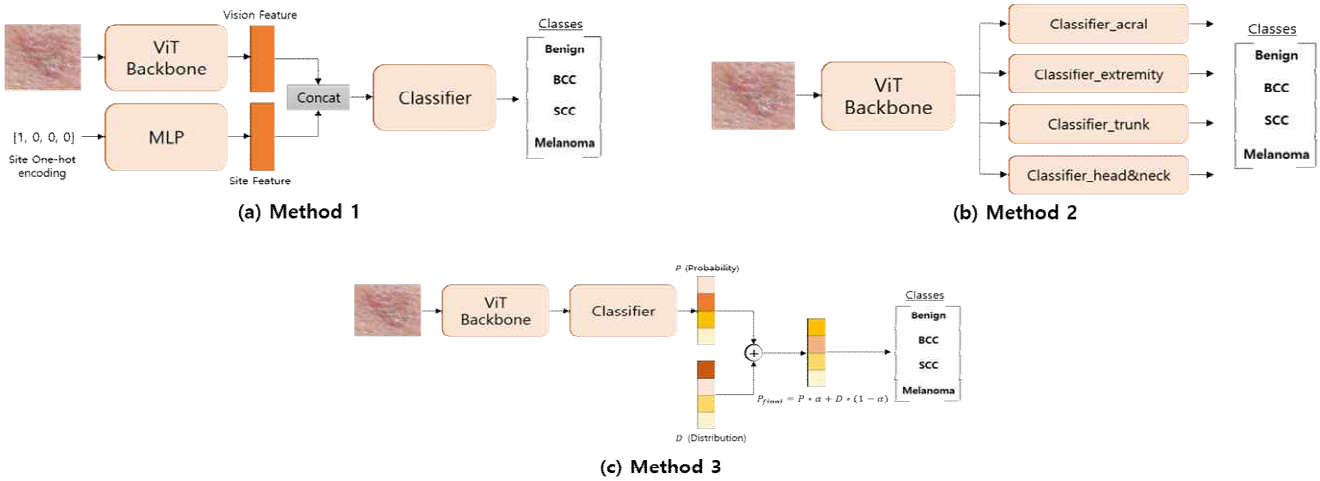

Fig. 2 Simple diagrams of three methods of site information fusion

2.2 Site Information Fusion

We propose three methods to effectively utilize anatomical site information for skin cancer classification. The first method involves creating site embedding through one-hot encoding of the anatomical site data, which is then fused with the visual features extracted by the existing ViT to perform the final classification, as illustrated in Fig. 2 (a). The second method involves first training the ViT and then creating separate classifiers for each site. In this approach, the trained ViT is kept frozen while the site-specific classifiers are retrained, as illustrated in Fig. 2 (b). The third method involves extracting the class distribution for each site from the training data and incorporates it into the training process. The probabilities generated by the ViT are combined with the class distributions for each site using an alpha value, which is a hyper-parameter, to make the final classification. This method is illustrated in Fig. 2 (c).

Table. 1 Accuracy of skin cancer classification based on each method and test data set

	ClinicalTrial	Demonstration
Baseline(ViT)	0.71	0.78
Method 1	0.72	0.77
Method 2	0.74	0.8
Method 3 (alpha=0.5)	0.75	0.8

2.3 Results

We conducted experiments on four models: the Baseline (ViT) model and three additional models that integrate site information in different ways. These experiments were performed using two test sets, ClinicalTrial and Demonstration. Overall, the results showed that utilizing site information led to slight performance improvements compared to the Baseline model. Notably, Method 3, which leverages the class distribution by site, achieved the highest performance on both test sets. For the experiment with Method 3, we conducted a total of nine tests by varying the alpha value from 0.1 to 0.9 in increments of 0.1. The highest performance was observed when alpha was set to 0.5. This suggests that both probability and distribution hold significant importance.

Ⅲ. Conclusion

In this study, we explored the integration of anatomical site information into a ViT-based model for skin cancer classification. Our experiments demonstrated that incorporating site-specific data generally improved model performance compared to the baseline ViT model. Among the methods tested, Method 3, which utilized the class distribution by site, consistently achieved the highest accuracy across different test sets. These findings highlight the value of combining clinical metadata with advanced machine learning techniques, suggesting a promising direction for enhancing automated skin cancer diagnosis.

ACKNOWLEDGMENT

This research was partly supported by the MSIT(Ministry of Science and ICT), Korea, under the ITRC(Information Technology Research Center) support program(IITP-2024-2020-0-01808) supervised by the IITP(Institute of Information & Communications Technology Planning & Evaluation) and National IT Industry Promotion Agency(NIPA) grant funded by the Korea government(MSIT) (S0252-21-1001, Development of AI Precision Medical Solution(Doctor Answer 2.0)).

References

[1] Garbe, Claus, and Ulrike Leiter. "Melanoma epidemiology and trends." *Clinics in dermatology* 27.1 (2009): 3-9.

[2] Apalla, Zoe, et al. "Epidemiological trends in skin cancer." *Dermatology practical & conceptual* 7.2 (2017): 1.

[3] DOSOVITSKIY, Alexey. "An image is worth 16x16 words: Transformers for image recognition at scale." *arXiv preprint arXiv:2010.11929* (2020)