Squamous Cell Carcinoma Margin Classification Using Vision Transformers from Digital Histopathology Images

So-Yun Park
IT Convergence Engineering
Kumoh National Institute of Technology
Gumi, Republic of Korea
alice3913@kumoh.ac.kr

Gelan Ayana
Medical IT Convergence Engineering
Kumoh National Institute of Technology
Gumi, Republic of Korea
gelan@kumoh.ac.kr

Se-woon Choe
IT Convergence Engineering
Kumoh National Institute of Technology
Gumi, Republic of Korea
sewoon@kumoh.ac.kr

***Abstract*— Skin cancer is one of the most prevalent cancers in the world, requiring the examination of tissue samples from all margins under a microscope through biopsy to determine whether cancer cells exist. Among skin cancers, squamous cell carcinoma (SCC) accounts for the majority of metastatic cancers and deaths related to non-melanoma skin cancers (NMSC). This study is the first to use the Vision Transformer (ViT) model to evaluate the margins obtained through biopsy. Various ViT model variants, including ViT-B16, ViT-B32, and ViT-L32, were experimented by adding additional layers. The results showed that the ViT-B16 model, when supplemented with additional layers, exhibited the best performance in classifying SCC images, achieving a high AUC value and a low loss value. This study also compared the performance of the ViT-B16 model with a CNN-based model, ResNet50, and found that the ViT-B16 significantly outperformed ResNet50. Specifically, the ViT-B16 model achieved an accuracy of 0.906 and an AUC of 0.905 in SCC margin classification, demonstrating its superiority over CNN-based models. This highlights the potential of the ViT model to surpass conventional CNNs, particularly showcasing its strength in handling complex medical image classification tasks.**

Keywords*— *squamous cell carcinoma, vision transformer, classification

I. INTRODUCTION

Skin cancer is a malignancy that arises from the skin, the outermost layer of the human body, and results from the abnormal proliferation of skin cells. Globally, one in three cancer diagnoses is skin cancer, with over 3.5 million incidences reported annually only in the USA [1]-[2]. Skin cancer can be broadly categorized into non-melanoma skin cancer (NMSC) and melanoma (MSC). NMSC is categorized into squamous cell carcinoma (SCC) and basal cell carcinoma (BCC). NMSC is the most common malignancy among Caucasians, with its incidence rate continuing to rise [2]. Notably, SCC accounts for a significant portion of metastatic cancers and deaths associated with NMSC [3].

SCC is classified into three differentiation stages, with well-differentiated SCC being Grade I and poorly differentiated SCC being Grades II and III. Grade IV SCC is classified as undifferentiated or invasive [4]. In this study, we focus on the diagnosis of tumor margins and consider all grades of SCC as malignant, while margins without cancer are assessed as normal or benign [3].

SCC is primarily diagnosed through dermoscopic examination or tissue biopsy, followed by Mohs micrographic surgery (MMS) [3]. After biopsy, tissue samples are taken from all margins to be examined microscopically to confirm whether cancer cells have been successfully removed. This method of margin assessment also plays a crucial role in ensuring the complete removal of tumor cells during treatment. However, microscopic examination relies heavily on the pathologist's judgment, requiring extensive experience and expertise to ensure accurate diagnosis [5].

The quality of histopathological images significantly impacts the accuracy of AI-based diagnostic systems, with various factors such as microscope quality, staining techniques, laboratory conditions, and reagent reliability playing a crucial role [6–8]. However, in resource-constrained settings, limited access to high-quality equipment and skilled personnel often results in low-quality images, which can hinder accurate diagnosis and treatment planning [9–12].

Over the past decade, convolutional neural networks (CNNs) have been widely employed for histopathological image analysis. However, their performance tends to degrade significantly when dealing with low-quality images [13–16]. To overcome this challenge, this study represents the first attempt to classify SCC margins using the Vision Transformer (ViT). ViT is rapidly emerging as a leading deep learning approach in the field of image processing. Unlike traditional CNNs, ViT divides the image into patches and processes each patch through a Transformer architecture to analyze the entire image [17]. This characteristic of ViT offers the potential to effectively analyze the complex patterns of skin tissues, aiming to achieve higher accuracy and efficiency compared to existing methods [18]-[22].

In this study, we propose a ViT-based model tailored to classify SCC margins in low-quality histopathological images. By evaluating the efficacy of ViT in this context, we aim to demonstrate its potential to improve diagnostic accuracy and clinical decision-making, even in resource-limited environments.

II. METHODS

A. Dataset and Preprocessing

In this study, tissue samples obtained from Jimma University Medical Center were used (data can be accessed at https://osf.io/3ma4p/; last accessed August 28, 2024) [3]. The collected dataset is composed of tissue sample images from patients, collected from various skin sites. Hematoxylin and eosin (H&E) staining was performed on all samples collected. This dataset comprises slides extracted from a total of 50 patients, of which 17 had well-differentiated SCC, 15 had moderately differentiated SCC, and 18 had invasive SCC. In this study, 345 normal tissue sample images were classified as margin-negative (normal images), while 483 images containing tumors were designated as margin-positive (tumor images) [3]. Example images are provided in Fig. 1. 90% of the entire dataset were used for training, and the remaining 10% were used for testing.

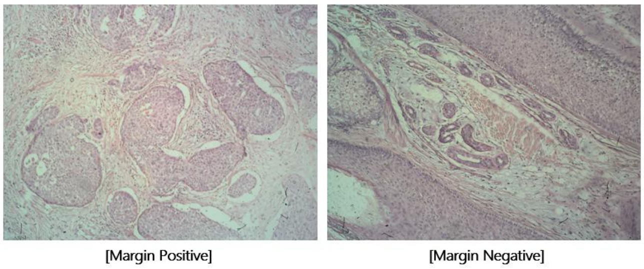

Fig. 1. Positive and negative images of SCC margin cells

The original resolution of the SCC margin cell images was 2048x1536 pixels, and all images were resized to 224x224 pixels. This is the preferred size for generating patches from the input images. Additionally, various techniques for augmentation such flipping, scaling, and rotation were used for data augmentation. To evaluate the generalization performance of the model and prevent overfitting, k-fold cross-validation was conducted.

B. *Proposed Model*

Throughout this research, a ViT-based training approach was employed to classify SCC cells as positive or negative. Consequently, a pre-trained model on the ImageNet dataset was used. The updated ViT architecture proposed in [23]-[27] was adopted. This structure has demonstrated superior performance in previous studies, particularly showing higher accuracy and stability compared to conventional models in image classification tasks. This architecture is designed by sequentially adding a flattening layer, batch normalization, and Dense layers after the MLP head of the ViT model, and the proposed structure can be seen in Fig. 2.

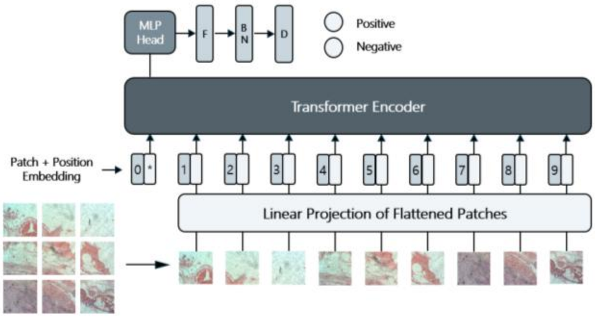
MLP Head
F
B N
D
Positive
Negative
Transformer Encoder
Patch + Position Embedding
0 *
1
2
3
4
5
6
7
8
9
Linear Projection of Flattened Patches

Fig. 2. Structure of ViT model with added layers

C. *Training Procedure*

The ViT model uses the Adam optimizer, known for its ability to handle large-scale datasets and high-dimensional parameter spaces effectively. Training was conducted using a learning rate of 0.0001, as determined optimal in prior research. The training process was performed on a high-performance computing cluster equipped with NVIDIA GPU 3090, which enabled efficient processing of computationally intensive tasks. To verify the performance improvement of the proposed architecture, training was performed using a total of 50 epochs. A batch size of 64 was used to balance memory efficiency and learning speed. Additionally, L2 regularization in the form of weight decay was applied to prevent overfitting by discouraging the learning of overly complex patterns. This settings also aimed to evaluate whether the proposed model could achieve comparable results under varying datasets and experimental conditions.

III. Results

This work conducted experiments based on the ViT model. The models used were ViT-B16, ViT-B32, and ViT-L32, and their performance was compared. As shown in TABLE I, the ViT-B16 model, with the proposed additional layers, outperformed the other models in classifying SCC images. The ViT-B16 achieved an accuracy of 0.906, an AUC of 0.905, and a loss value of 0.402, surpassing ViT-B32 in all aspects. ViT-L16 had a loss value of 0.448, confirming that ViT-B16 was the best-performing model. Additionally, as illustrated in Fig. 3, ViT-B16 exhibited relatively stable convergence, and the ROC curve, with a value of 0.905, demonstrated the strong discriminatory power of the model. The confusion matrix further supported the accuracy of ViT-B16, showing very few misclassification cases.

TABLE I: EXPERIMENTAL RESULTS AND PERFORMED COMPARISONS FOR DIFFERENT MODELS

Model	**Acc.**	**Prec.**	**Recall**	**F-1 score**	**AUC**	**Loss**
ViT-B16	0.906	0.902	0.906	0.902	0.905	0.402
ViT-B32	0.878	0.88	0.884	0.878	0.882	0.447
ViT-L16	0.904	0.902	0.904	0.902	0.902	0.448

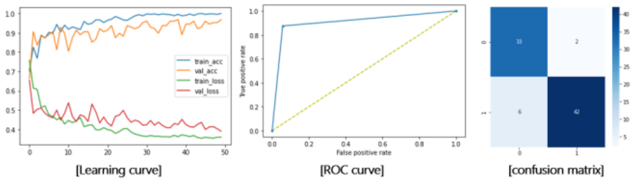

Fig. 3. Visualization the results of the ViT-B16 model as learning curve, ROC curve, and confusion matrix

Additionally, as shown in TABLE II, modifying the ViT-B16 model with different layers resulted in lower performance compared to the model with the proposed layers, confirming that the suggested layers contributed to the improvement in performance.

TABLE II: Experimental results and comparisons for different layer additions were performed : *F*, flatten; *B*, batch normalization; *D*, Dense

Layer	**Acc.**	**Prec.**	**Recall**	**F-1 score**	**AUC**
ViT+F+B+D+B+D	0.884	0.882	0.886	0.882	0.885
ViT+B+D+F+B+D	0.902	0.902	0.898	0.898	0.897

Lastly, as presented in TABLE III, the comparison with ResNet50, a traditional CNN model, demonstrates that the Vision Transformer model can achieve superior performance in image classification tasks.

TABLE III: Comparisons with experimental results for CNNs were performed.

Model	**Acc.**	**Prec.**	**Recall**	**F-1 score**	**AUC**	**Loss**
ViT-B16	0.906	0.902	0.906	0.902	0.905	0.402
ResNet50	0.816	0.79	0.81	0.76	0.762	0.538

IV. Discussion

In this study, the ViT was used to classify histopathological margin images of SCC. Notably, the ViT-B16 model, with additional layers, demonstrated excellent performance in SCC margin image classification tasks. The ViT-B16 model achieved a high accuracy of 0.906 and an AUC value of 0.905, outperforming other ViT variants (ViT-B32, ViT-L32) as well as the traditional CNN model

(ResNet50). It also recorded a lower loss value of 0.402 compared to other models. This indicates that the structural characteristics of the ViT-B16 model and the design of the additional layers significantly enhanced the model's learning ability and classification performance.

In experiments where different layers were applied to the ViT-B16 model, performance actually decreased, confirming that the originally proposed layer structure was the most suitable for achieving optimal performance. These results suggest that careful adjustments to the model's architecture can impact the performance of the ViT model.

Furthermore, the superior performance of the ViT model compared to traditional CNN models like ResNet50 highlights that the modeling approach of Vision Transformers—processing images based on patches—can more effectively learn fine image features that traditional CNN models may struggle to capture. These findings suggest the potential for Vision Transformer models to be improvement over CNNs in a variety of image analysis tasks.

This study has limitations. First, this study considered limited number of parameters that affect performance of vision transformers. Future studies should consider experimenting with the various types deep learning parameters to improve performance. Second, in this study, only three variants of additional layers were experimented, further experiment with more additional layers might improve performance. Lastly, this study was conducted using data from a single institution, and the proposed method was not validated on independent external datasets. Future studies should aim to assess the generalizability and robustness of the proposed approach across diverse datasets and clinical environments.

V. Conclusion

Globally, skin cancer is commonly diagnosed through Mohs micrographic surgery for tumor removal, followed by histopathological analysis of margin samples to confirm the complete excision of cancer cells. Margin evaluation plays a crucial role in ensuring the thorough removal of tumor cells, and thus, this study focused on classifying SCC margins using the ViT.

Several variants of the ViT model (ViT-B16, ViT-B32, ViT-L32) were used to assess performance. The results showed that the ViT-B16 model, with additional layers proposed in previous studies, demonstrated the best performance, even surpassing the classification ability of ResNet50, a traditional CNN model. The ViT-B16 model achieved an accuracy of 0.906 and an AUC value of 0.905, effectively classifying SCC margin images. These findings suggest that the ViT model perform better than CNNs for SCC images classification.

Future research should focus on further optimizing the performance of ViT models. First, it is necessary to analyze the training process by applying various parameters that maximizes the generalization performance of the model. Second, to further enhance the performance of the ViT-B16 model, experimenting with different layer configurations beyond the existing ones to find the best combination is essential. Specifically, it is important to explore how introducing additional attention mechanisms or more complex network architectures could improve the model's performance. Furthermore, to expand the applicability of ViT models, it is critical to validate their performance across diverse medical image datasets and evaluate whether the model consistently achieves high performance in image classification tasks across various medical fields. Such research could open up new possibilities for the widespread use of ViT models in a range of medical image analysis applications, including skin cancer diagnosis.

Acknowledgment

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (RS-2023-00240521).

References

[1] H.S. Balaha, and A.E.S. Hassan, “Skin cancer diagnosis based on deep transfer learning and sparrow search algorithm”, *Neural Computing and Applications*, vol. 35, pp. 815–853, Sep. 2023, doi: 10.1007/s00521-022-07762-9

[2] Roky, Amdad Hossain, et al. "Overview of skin cancer types and prevalence rates across continents.", *Cancer Pathogenesis and Therapy*, vol. 2, pp. E01-E36, Aug. 2024, doi: 10.1016/j.cpt.2024.08.002

[3] Wako, Beshatu Debela, et al. "Squamous cell carcinoma of skin cancer margin classification from digital histopathology images using deep learning.", *Cancer Control*, vol. 29, pp. 1-16, Oct. 2022, doi: 10.1177/10732748221132528

[4] Koyuncuer. Ali. "Histopathological evaluation of non-melanoma skin cancer", *World journal of surgical oncology*, vol.12, pp. 1-6, May. 2014, doi: 10.1186/1477-7819-12-159

[5] S. Jiang, H. Li and Z. Jin, "A Visually Interpretable Deep Learning Framework for Histopathological Image-Based Skin Cancer Diagnosisr", *IEEE Journal of Biomedical and Health Informatics*, vol.25, pp. 1483-1494, Jan. 2021, doi: 10.1109/JBHI.2021.3052044

[6] Baxi, V.; Edwards, R.; Montalto, M.; Saha, S. Digital Pathology and Artificial Intelligence in Translational Medicine and Clinical Practice. Mod. Pathol. 2022, 35, 23–32, doi:10.1038/s41379-021-00919-2.

[7] Meyerholz, D.K.; Beck, A.P. Principles and Approaches for Reproducible Scoring of Tissue Stains in Research. Lab. Investig. 2018, 98, 844–855, doi:10.1038/s41374-018-0057-0.

[8] Kreiss, L.; Jiang, S.; Li, X.; Xu, S.; Zhou, K.C.; Lee, K.C.; Mühlberg, A.; Kim, K.; Chaware, A.; Ando, M.; et al. Digital Staining in Optical Microscopy Using Deep Learning - a Review. *PhotoniX* **2023**, *4*, 34, doi:10.1186/s43074-023-00113-4.

[9] Bai, B.; Yang, X.; Li, Y.; Zhang, Y.; Pillar, N.; Ozcan, A. Deep Learning-Enabled Virtual Histological Staining of Biological Samples. *Light Sci. Appl.* **2023**, *12*, 57, doi:10.1038/s41377-023-01104-7.

[10] Haghighat, M.; Browning, L.; Sirinukunwattana, K.; Malacrino, S.; Khalid Alham, N.; Colling, R.; Cui, Y.; Rakha, E.; Hamdy, F.C.; Verrill, C.; et al. Automated Quality Assessment of Large Digitised Histology Cohorts by Artificial Intelligence. Sci. Rep. 2022, 12, 5002, doi:10.1038/s41598-022-08351-5.

[11] RAHMAN, T.Y.; MAHANTA, L.B.; CHAKRABORTY, C.; DAS, A.K.; SARMA, J.D. Textural Pattern Classification for Oral Squamous Cell Carcinoma. *J. Microsc.* 2018, *269*, 85–93, doi:10.1111/jmi.12611.

[12] Wako, B.D.; Dese, K.; Ulfata, R.E.; Nigatu, T.A.; Turunbedu, S.K.; Kwa, T. Squamous Cell Carcinoma of Skin Cancer Margin Classification From Digital Histopathology Images Using Deep Learning. Cancer Control 2022, 29, 107327482211325, doi:10.1177/10732748221132528.

[13] Ahmed, A.A.; Abouzid, M.; Kaczmarek, E. Deep Learning Approaches in Histopathology. Cancers (Basel). 2022, 14, 5264, doi:10.3390/cancers14215264.

[14] Jeong, H.K.; Park, C.; Jiang, S.W.; Nicholas, M.; Chen, S.; Henao, R.; Kheterpal, M. Image Quality Assessment Using Convolutional Neural Network in Clinical Skin Images. JID Innov. 2024, 4, 100285, doi:10.1016/j.xjidi.2024.100285.

[15] Tang, S.; Jing, C.; Jiang, Y.; Yang, K.; Huang, Z.; Wu, H.; Cui, C.; Shi, S.; Ye, X.; Tian, H.; et al. The Effect of Image Resolution on Convolutional Neural Networks in Breast Ultrasound. Heliyon 2023, 9, e19253, doi:10.1016/j.heliyon.2023.e19253.

[16] Hu, C.; Sapkota, B.B.; Thomasson, J.A.; Bagavathiannan, M. V. Influence of Image Quality and Light Consistency on the Performance of Convolutional Neural Networks for Weed Mapping. Remote Sens. 2021, 13, 2140, doi:10.3390/rs13112140.

[17] A. Vaswani, et al. "Attention Is All You Need", In Proceedings of the 31st International Conference on Neural Information Processing Systems, pp. 6000-6010, Dec. 2017, doi: 10.48550/arXiv.1706.03762

[18] Atabansi, C.C.; Nie, J.; Liu, H.; Song, Q.; Yan, L.; Zhou, X. A Survey of Transformer Applications for Histopathological Image Analysis: New Developments and Future Directions. Biomed. Eng. Online 2023, 22, 96, doi:10.1186/s12938-023-01157-0.

[19] Maurício, J.; Domingues, I.; Bernardino, J. Comparing Vision Transformers and Convolutional Neural Networks for Image Classification: A Literature Review. Appl. Sci. 2023, 13, 5521, doi:10.3390/app13095521.

[20] Ayana, G.; Barki, H.; Choe, S. Pathological Insights: Enhanced Vision Transformers for the Early Detection of Colorectal Cancer. Cancers (Basel). 2024, 16, 1441, doi:10.3390/cancers16071441.

[21] Raghu, M.; Unterthiner, T.; Kornblith, S.; Zhang, C.; Dosovitskiy, A. Do Vision Transformers See Like Convolutional Neural Networks? 2021.

[22] Ayana, G.; Lee, E.; Choe, S. Vision Transformers for Breast Cancer Human Epidermal Growth Factor Receptor 2 Expression Staging without Immunohistochemical Staining. Am. J. Pathol. 2023, 194, 402–414, doi:10.1016/j.ajpath.2023.11.015.

[23] G. Ayana, H. Barki, and S.W. Choe, “Pathological Insights: Enhanced Vision Transformers for the Early Detection of Colorectal Cancer”, *Cancers*, vol.16(7), Apr. 2024, doi: 10.3390/cancers16071441

[24] G.Ayana, and S.W. Choe, “Vision Transformers-Based Transfer Learning for Breast Mass Classification From Multiple Diagnostic Modalities”, *Journal of Electrical Engineering & Technology*, vol.19, pp. 3391-3410, Apr. 2024, doi: 10.1007/s42835-024-01904-w

[25] G.Ayana, E.J. Lee, and S.W. Choe, “Vision Transformers for Breast Cancer Human Epidermal Growth Factor Receptor 2 Expression Staging without Immunohistochemical Staining”, *The American Journal of Pathology*, vol. 194, pp.402-414, Mar. 2024, doi: 10.1016/j.ajpath.2023.11.015

[26] G.Ayana, et al. “Vision-Transformer-Based Transfer Learning for Mammogram Classification”, *Diagnostics*, vol.13(2), Jan. 2023, doi: 10.3390/diagnostics13020178

[27] G.Ayana, and S.W. Choe, “BUViTNet: Breast Ultrasound Detection via Vision Transformers”, *Diagnostics*, vol.12(11), Nov. 2022, doi: 10.3390/diagnostics12112654