

A spatial-spectral vision transformer model for head and neck cancer detection with hyperspectral, RGB, and synthesized RGB histologic images

Hemanth Pasupuleti ^{a,b}, Ling Ma ^{a,b}, Xiaohu Guo ^c, Baowei Fei ^{a,b,d,*}

^a Center for Imaging and Surgical Innovation, University of Texas at Dallas, Richardson, TX

^b Department of Bioengineering, University of Texas at Dallas, Richardson, TX

^c Department of Computer Science, University of Texas at Dallas, Richardson, TX

^d Department of Radiology, University of Texas Southwestern Medical Center, Dallas, TX

*Corresponding author: bfei@utdallas.edu, Website: <https://fei-lab.org>

ABSTRACT

The potential benefit of employing hyperspectral imaging (HSI) over the normal color images, *i.e.*, RGB, has been explored to improve cancer detection. In this study, we evaluate the efficacy of deep learning models in detecting head and neck squamous cell carcinoma (SCC) in histological images using HSI. We used 77 whole histologic slides from 51 patients to train DinoV2 vision transformer model and a ResNet-152 model on RGB, HSI, and HSI-synthesized RGB images, respectively. A spatial-spectral vision transformer (SST) model with spectral attention was also introduced for comparison and evaluation. Our study resulted in three major findings. First, we found that the SST model trained on HSI performed the best over other models with 79% accuracy, 74.37% specificity, and 78.92% sensitivity. Second, models trained using RGB data suffered from severe imbalance between sensitivity and specificity. Finally, the DinoV2 model trained on HSI was found to have significantly higher (10%) sensitivity compared to its RGB alternative. The proposed method of using the spatial-spectral vision transformer model shows the advantage of hyperspectral image in terms of sensitivity, accuracy, and computation for detecting squamous cell carcinoma in histologic slides.

Keywords: Hyperspectral imaging (HSI), head and neck squamous cell carcinoma, histology, spatial-spectral transformer,

1. INTRODUCTION

Head and neck cancer, being the seventh most common cancer globally, is one of the leading causes of mortality worldwide with 325,000 deaths annually ¹, necessitating continuous advancements in diagnostic treatment. Squamous cell carcinoma (SCC) is one of the most common forms of head and neck cancer with 90 percent of cancers that occur in the head and neck. Medical studies indicate that early and accurate detection is critical for improving patient outcomes, guiding treatment decisions, and enhancing survival rate ². Several deep learning models for detecting SCC from histologic slides have been investigated using digitized histological images. Halicek *et al* used a convolutional neural network (CNN) based on the Inception-v4 architecture to detect head and neck cancer in digitalized whole slide image (WSI) of hematoxylin and eosin (H&E)-stained histological slides ^{3,4}. They performed patch wise prediction and aggregation for the whole slide to predict the final output, achieving an AUC of 0.92 for SCC on a dataset of 97 patients. Mavuduru *et al* ⁵ adapted a U-Net architecture to detect and segment SCC in digitalized whole slide images, achieving AUC of 0.89.

Hyperspectral imaging (HSI) is an optical imaging technique that is emerging in medical applications for disease diagnosis ⁶⁻⁸. Unlike normal color imaging that only has three channels (RGB), HSI captures images of multiple bands of wavelength thus outputting a three-dimensional cube with two spatial dimensions and one spectral dimension. Wei *et al* ⁹ utilized unsupervised feature extraction using principal component analysis (PCA) to perform classification. Huang *et al* ¹⁰ utilized PCA and Gabor wavelet features on HSI images to train CNN for classifying blood cells. Wang *et al* ¹¹ combined 3D-CNN and 3D attention to use spatial and spectral features for leukocyte classification. Although many studies highlight superior performance of HSI over RGB, they did not explore distinct benefits of using HSI in detail. In this preliminary study, we investigate the performance benefit of deep learning models using HSI over RGB images in detecting SCC from WSI regarding the usefulness of spectral information and the improvement that HSI could bring to certain cancer detection performance.

Traditional vision transformer models¹² trained on RGB apply convolution layer to embed the channel information into the patch followed by linear layer to get the patch embedding, while the ResNet model¹³ performs 2D convolutions to extract information from the channel dimension. Although these models perform extremely well classifying RGB images, when applied to hyperspectral images, the spectral dimension is often neglected or may not be fully utilized. Hence, we introduced a spatial-spectral vision transformer model adapted from Scheibenreif et al¹⁴ to apply the attention mechanism on both spatial and spectral dimensions. This attention gives priority to both spatial information and spectral information.

2. METHODS

2.1 Hyperspectral images and data preparation

H&E-stained histological slides were scanned using our custom hyperspectral imaging system¹⁵, which was capable of autofocusing and whole-slide imaging with a snapscan hyperspectral camera^{16,17}. Considering the large file size and long acquisition time of acquiring high-resolution hyperspectral images, we chose to acquire low-resolution whole-slide hyperspectral histological images at an objective magnification of 4×. Each image had 87 spectral bands covering the wavelength range of 460-720 nm. Digitized RGB whole-slide histological images were obtained from previous work^{3,4} at 40× objective magnification. Then, the hyperspectral and RGB images were automatically aligned and cropped into patches using a standardized automated pipeline, and we used an unsupervised super-resolution method^{16,17} to reconstruct high-resolution HSI images using the high-resolution RGB and low-resolution HSI image pairs. In addition to HSI and RGB, we also investigated the classification performance using HSI-synthesized RGB (syn RGB). To generate pseudo RGB images, we initially trained a simple U-Net model^{18,19} and modified the input to accept HSI image.

Our dataset in this work consisted of 52 patients with either tumor or normal whole-slide images, where 30 patients were used for training, 10 patients for validation, and 11 patients for testing. This resulted in a training dataset with 34044 tumor and 35887 non-tumor patches, a validation set with 18452 tumor and 23747 non-tumor patches, and a test set with 19544 tumor and 21853 non-tumor patches. A total of 77 whole slides with 153527 patches were utilized for this dataset. Figure 1. shows a sample patch from a whole slide that was labelled as tumor slide.

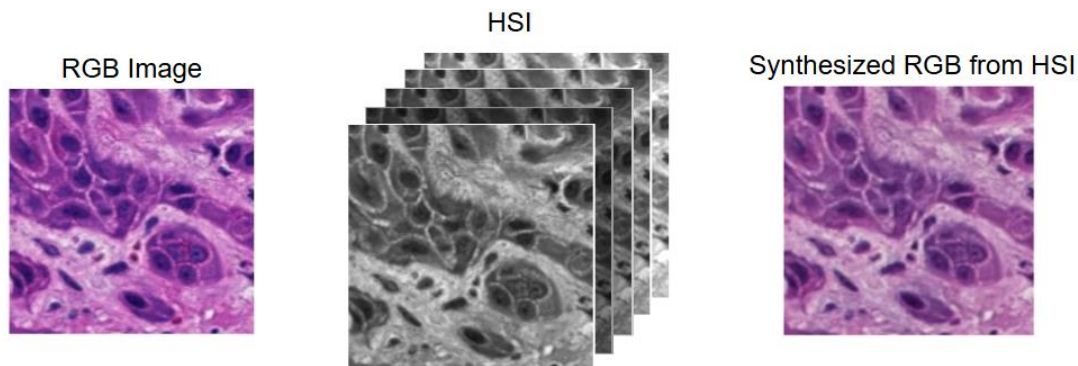


Figure 1. Sample tumor patch from training dataset consisting of RGB image, HSI, and HSI-synthesized RGB image.

2.2 Spatial-Spectral Vision Transformer

Figure 2 shows the architecture of the spatial-spectral vision transformer model. Initially, the hyperspectral cubes were split into blocks where the height and width of the block correspond to a spatial patch size P_{spa} and the depth of the block was the spectral patch size P_{spe} . Since the input cube was of dimension $224 \times 224 \times 87$, we split the cube using $P_{spa} = 16$ and into 5684 blocks of dimension $16 \times 16 \times 3$ each. Each block was embedded with its own patch embedding using linear layer per block. These embedded blocks were sent into spatial transformer where transformer attention was applied on spatial dimension. The outputs are reshaped and sent into a spectral transformer where attention was applied on spectral dimension, considering the spatial dimension as a batch sample. After multiple layers of spatial and spectral transformers, feature aggregation was applied across all blocks, followed by a linear layer to predict the output class.

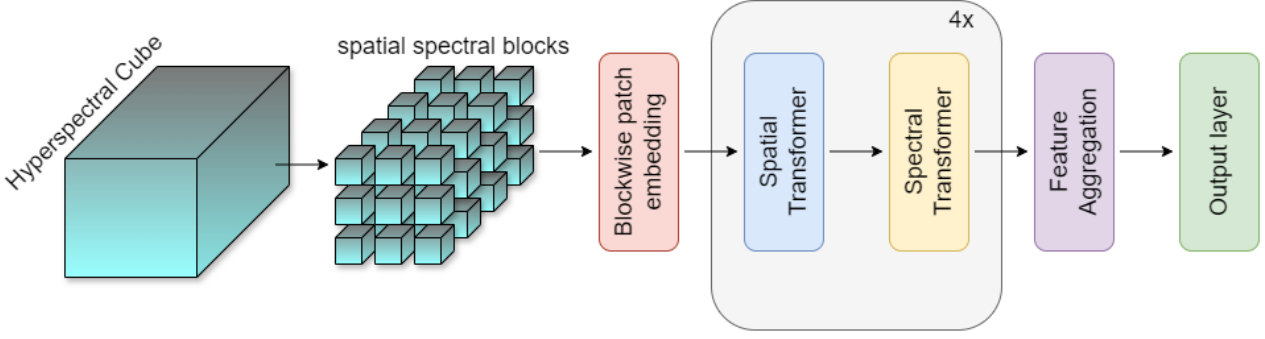


Figure 2. Architecture of spatial-spectral vision transformer for hyperspectral data.

2.3 Model Training

The reconstructed high-resolution HSI patches were used to train multiple models along-side with high-resolution RGB patches. All training patches were spatially resized from 200×200 pixels to 224×224 pixels, augmented with horizontal and vertical flips, and normalized. RGB images were normalized using ImageNet²⁰ normalization while HSI images were normalized by calculating the channel-wise mean and standard deviation over the training images. We trained and evaluated ResNet-152¹³, DinoV2 vision transformer¹², and Spatial-Spectral Vision Transformer models on RGB images, HSI images, and HSI-synthesized RGB, which were of the same image resolution. For a fair comparison, all models are trained with no pretrained weights. We also introduced a spatial-spectral attention-based vision transformer as discussed in section 2.3 to evaluate if spectral attention can improve the performance. All models were trained and tested on a single NVIDIA RTX A6000 with 48 GB of GPU memory with a batch size of 64 except for spatial-spectral transformer model, which was trained at 32 batch size to fit the memory. We train all models using AdamW optimizer²¹ for 10 epochs with a learning rate starting at 0.0001 and reducing every 2 epochs up on plateau with a factor of 0.8. The loss function used for this task was cross entropy loss and we selected the model performing the best on the validation set for testing.

2.4 Evaluation Metrics

In this study, we evaluated the models based on accuracy, f1-score, sensitivity, specificity, and memory and computation costs. Accuracy (Eq. 1) is defined as the ratio of number of patches predicted correctly to the total number of patches in the set. F1-score (Eq. 2) measures the harmonic mean of precision and recall, where precision is ratio of number of patches predicted correctly for a label to total number of patches the model predicted with the label, and recall is ratio of number of patches predicted correctly to the actual number of patches for the label. Sensitivity (Eq. 3) is another name for recall, while specificity (Eq. 4) is the ratio of number of patches that the model predicted negative to the total number of patches that are negative. We define the equations below with true positive (TP), true negative (TN), false positive (FP), false negative (FN), where positive represents cancerous patch, and negative represents non-cancerous patch. Further, to evaluate the memory required for the model, we report the total number of parameters for the model, and computation cost is represented by number of floating-point operations (FLOPs) being performed in the model.

$$\text{Overall Accuracy} = \frac{TP + TN}{TP + FP + TN + FN} \quad (1)$$

$$\text{F1 Score} = \frac{TP}{TP + \frac{1}{2}(FP + FN)} \quad (2)$$

$$\text{Sensitivity} = \frac{TP}{TP + FN} \quad (3)$$

$$\text{Specificity} = \frac{TN}{TN + FP} \quad (4)$$

3. RESULTS

Table 1 shows the model performances of ResNet-152 and DinoV2 on three types of data: RGB images, reconstructed high-resolution HSI data, and HSI-synthesized RGB images. The spatial-spectral transformer was trained and evaluated

on HSI data only to enforce spectral attention mechanism on the spectral dimension. The results indicate that spatial-spectral transformer achieved the highest scores with 79% accuracy, 79% F1 score, and 78.92% sensitivity.

In comparison, the DinoV2 model trained on HSI data also demonstrated a strong performance, achieving 78% accuracy, 78% f1-score and 74.13% sensitivity, showing higher sensitivity and F1-scores than both ResNet and DinoV2 models trained on RGB, and HSI-synthesized RGB images. However, while the DinoV2 model trained on RGB images achieved highest specificity at 86.56%, it also revealed a notable drawback across all data types (RGB, HSI, or syn RGB). Specifically, both ResNet and DinoV2 vision transformer that focuses on spatial dimensions, exhibited a significant imbalance between sensitivity and specificity, where high specificity came with a cost of reduced sensitivity.

In contrast, the spatial-spectral transformer (SST) effectively addressed this imbalance by incorporating both spatial and spectral attention mechanisms, resulting in a more balanced sensitivity and specificity scores. This model not only delivered superior performance, but it also stood out in terms of having the least memory and computational requirements with only 1.68 million parameters and 11.49 million floating-point operations. Despite SST being 300× smaller than DinoV2 model and requiring 100x less floating-point operations, it achieved 3% higher accuracy and 4% better in F1 score than the DinoV2 model when trained on RGB images. These findings highlight strong performance advantages of HIS over RGB models.

Table 1. Comparison of model performances on RGB data, HSI data, and RGB generated from HSI data on test dataset.

Model	Accuracy	F1-Score	Sensitivity	Specificity	# of params	FLOPs
ResNet-152 (RGB)	0.74	0.74	0.6766	0.7932	58M	11.33G
ResNet-152 (HSI)	0.74	0.74	0.6387	0.8312	58M	11.33G
ResNet-152 (syn RGB)	0.71	0.71	0.5905	0.8166	58M	11.33G
DinoV2 (RGB)	0.77	0.76	0.6415	0.8656	303M	1.36G
DinoV2 (HSI)	0.78	0.78	0.7413	0.8233	303M	1.36G
DinoV2 (syn RGB)	0.76	0.75	0.6586	0.8405	303M	1.36G
Spatial-Spectral Transformer (HSI)	0.79	0.79	0.7892	0.7437	1.68M	11.49M

4. DISCUSSION & CONCLUSION

In this study, we used histological images of head and neck tissue samples obtained via hyperspectral imaging (HSI) to compare the performance of HSI over RGB images in detecting cancerous patches from whole-slide images using deep learning models. We trained ResNet and Vision Transformer models on RGB images, HSI, and HSI-synthesized RGB images, while also introducing a spatial-spectral transformer model to enable spectral attention. The results demonstrated that models trained on HSI data showed higher sensitivity, with the spatial-spectral transformer achieving the highest average accuracy of 79%, an F1 score of 79%, and sensitivity of 78.92%, while maintaining only 1.68 million parameters and 11.49 million floating point operations.

The study revealed that the spatial-spectral model, when applied to HSI data, outperformed other models, showing a 10% improvement in detecting cancerous patches compared to RGB images. Moreover, the ResNet-152 and DinoV2 models trained on RGB and synthesized RGB images showed similar results, but they suffered from imbalances between sensitivity and specificity. Specifically, while the DinoV2 (RGB) model achieved the highest specificity of 86.56%, it lacked sufficient sensitivity. This imbalance was addressed by the spatial-spectral transformer model on the HSI data, which provided a better balance between sensitivity and specificity.

In conclusion, the proposed method using hyperspectral images, particularly with the spatial-spectral transformer model, shows significant potential in improving cancer detection, offering better performance than RGB images, while

maintaining balanced sensitivity and specificity. Future work will involve a deeper analysis of generalizability of HSI and RGB models with varying data distributions, increased dataset, and evaluation with pretraining.

ACKNOWLEDGMENTS

Research reported in this publication was supported in part by the National Cancer Institute of the National Institutes of Health under Award Number R01CA288379 and R01CA204254 and by the Cancer Prevention and Research Institute of Texas (CPRIT) under Award Number RP240289 and RP240542. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

REFERENCES

- [1] Gormley, M., Creaney, G., Schache, A., Ingarfield, K., and Conway, D. I., "Reviewing the epidemiology of head and neck cancer: definitions, trends and risk factors," *British Dental Journal*, 233(9), 780-786 (2022).
- [2] Marur, S., and Forastiere, A. A., "Head and Neck Squamous Cell Carcinoma: Update on Epidemiology, Diagnosis, and Treatment," *Mayo Clinic Proceedings*, 91(3), 386-396 (2016).
- [3] Halicek, M., Shahedi, M., Little, J. V., Chen, A. Y., Myers, L. L., Sumer, B. D., and Fei, B., "Head and Neck Cancer Detection in Digitized Whole-Slide Histology Using Convolutional Neural Networks," *Scientific Reports*, 9(1), 14043 (2019).
- [4] Halicek, M., Shahedi, M., Little, J. V., Chen, A. Y., Myers, L. L., Sumer, B. D., Fei, B., Tomaszewski, J. E., and Ward, A. D., "Detection of squamous cell carcinoma in digitized histological images from the head and neck using convolutional neural networks," *Proc. SPIE 10956, Medical Imaging 2019: Digital Pathology*, 109560K (2019).
- [5] Mavuduru, A., Halicek, M., Shahedi, M., Little, J., Chen, A., Myers, L., and Fei, B., "Using a 22-layer U-Net to perform segmentation of squamous cell carcinoma on digitized head and neck histological images," *Proc. SPIE 11320, Medical Imaging 2020: Digital Pathology*, 113200C (2020).
- [6] Lu, G., and Fei, B., "Medical hyperspectral imaging: a review," *J Biomed Opt*, 19(1), 10901 (2014).
- [7] Ma, L., and Fei, B., "Comprehensive review of surgical microscopes: technology development and medical applications," *Journal of Biomedical Optics*, 26(1), 010901 (2021).
- [8] Tran, M., Ma, L., Little, J., Chen, A., and Fei, B., "Thyroid carcinoma detection on whole histologic slides using hyperspectral imaging and deep learning," *Proc. SPIE 12039, Medical Imaging 2022: Digital and Computational Pathology*, 120390H (2022).
- [9] Wei, X., Li, W., Zhang, M., and Li, Q., "Medical Hyperspectral Image Classification Based on End-to-End Fusion Deep Neural Network," *IEEE Transactions on Instrumentation and Measurement*, 68(11), 4481-4492 (2019).
- [10] Huang, Q., Li, W., Zhang, B., Li, Q., Tao, R., and Lovell, N. H., "Blood Cell Classification Based on Hyperspectral Imaging With Modulated Gabor and CNN," *IEEE Journal of Biomedical and Health Informatics*, 24(1), 160-170 (2020).
- [11] Wang, Q., Wang, J., Zhou, M., Li, Q., Wen, Y., and Chu, J., "A 3D attention networks for classification of white blood cells from microscopy hyperspectral images," *Optics & Laser Technology*, 139, 106931 (2021).
- [12] Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., and El-Nouby, A., "Dinov2: Learning robust visual features without supervision," *arXiv preprint arXiv:2304.07193*, (2023).
- [13] He, K., Zhang, X., Ren, S., and Sun, J., "Deep Residual Learning for Image Recognition." *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770-778 (2016).
- [14] Scheibenreif, L., Mommert, M., and Borth, D., "Masked vision transformers for hyperspectral image classification," *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2166-2176 (2023).
- [15] Ma, L., Ha, A., Zainab, I., Rathgeb, A., Mubarak, H., and Fei, B., "An automatic processing framework for hyperspectral histologic images and benchmark dataset." *Proc. SPIE 13413, Medical Imaging 2025: Digital and Computational Pathology* (2025).
- [16] Ma, L., Rathgeb, A., Tran, M., and Fei, B., "Unsupervised Super Resolution Network for Hyperspectral Histologic Imaging," *Proc. SPIE 12039, Medical Imaging 2022: Digital and Computational Pathology*, 120390P (2022).

- [17] Ma, L., Rathgeb, A., Mubarak, H., Tran, M., and Fei, B., "Unsupervised super-resolution reconstruction of hyperspectral histology images for whole-slide imaging," *Journal of Biomedical Optics*, 27(5), 056502 (2022).
- [18] Ronneberger, O., Fischer, P., and Brox, T., "U-Net: Convolutional Networks for Biomedical Image Segmentation," *Medical image computing and computer-assisted intervention–MICCAI 2015*, 234-241 (2015).
- [19] Ma, L., Sherey, J., Palsgrove, D., and Fei, B., "Conditional generative adversarial network (cGAN) for synthesis of digital histologic images from hyperspectral images," *Proc. SPIE 12471, Medical Imaging 2023: Digital and Computational Pathology*, 124711D (2023).
- [20] Deng, J., Dong, W., Socher, R., Li, L. J., Kai, L., and Li, F.-F., "ImageNet: A large-scale hierarchical image database," *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 248-255 (2009).
- [21] Loshchilov, I., "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, (2017).