Evaluation of PLIP model performance using pathology images and notes based on OMOP-CDM

Harrin Kim[1], Min-Gyu Kim[2], Junhyuk Chang[1], Rae Woong Park[1,2]

[1]Department of Biomedical Sciences, Ajou University Graduate School of Medicine;

[2]Department of Biomedical Informatics, Ajou University School of Medicine

Background

One of the challenges in pathology analysis is the processing of high-resolution Whole Slide Images (WSI)[1]. The acquisition of high-quality digital images is crucial for accurate interpretation in pathology[2]. Additionally, the collection of large annotated datasets required for models presents significant difficulties.

The Pathology Language and Image Pre-Training (PLIP) model, trained on large-scale pathology image-text pairs, is able to provide more relevant feature extraction for medical image analysis[3]. PLIP is modeled after the Contrastive Language-Image Pretraining (CLIP) approach. The CLIP model, developed by OpenAI, was trained on a substantial dataset of 400 million image-text pairs collected from the internet[4]. However, there has not been a study evaluating PLIP with real-world data.

This study compares and evaluates the ability of the PLIP and CLIP models to identify organ tissues when provided with organ labels from OMOP-CDM notes and corresponding pathology images as inputs.

Methods

This study utilized pathology images and reports from Ajou University Medical Center (AUSOM) for the year 2021. We extracted patient data with reports from the note table in the OMOP-CDM. Pathology IDs, which are unique identifiers for slides, were used to link notes and images, enabling the integration of image data into the CDM. Organ labels from the notes were utilized. Images were sampled according to the distribution of these labels. The tissue sections of the images were cropped and resized to 224x224 pixels for use as inputs.

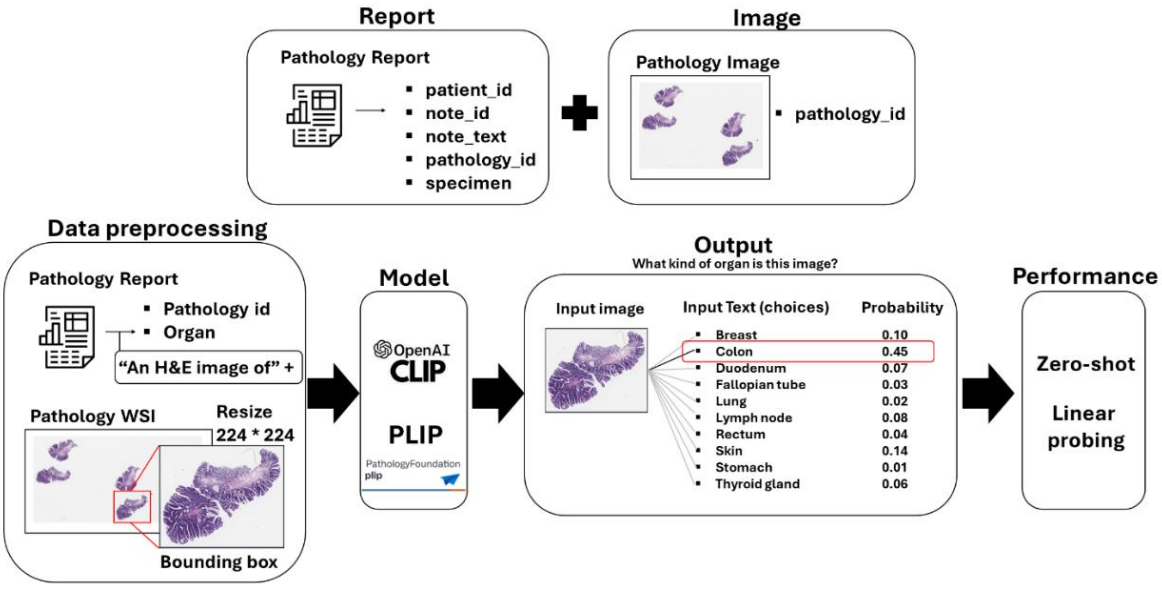

Figure 1. The workflow of a study framework.

Using CLIP and PLIP, we evaluated the models' ability to identify the organ from the tissue images. The evaluation was conducted using two approaches: zero-shot and linear probing. For linear probing, the dataset was split into 70% for training and 30% for testing. The accuracy of organ identification was measured using the F1 score, with the gold standard serving as the benchmark for comparison.

Results

We sampled 196 WSIs from a total of 73,231 whole slide images (WSIs) and cropped the tissue sections, resizing them to 224x224 pixels, resulting in a total of 1078 images. Among 274 labels, we selected the 12 most frequently occurring ones.

From the distribution of organ types, the most common organ observed was the stomach, followed by the colon, skin, breast, and lymph nodes (Figure 2).

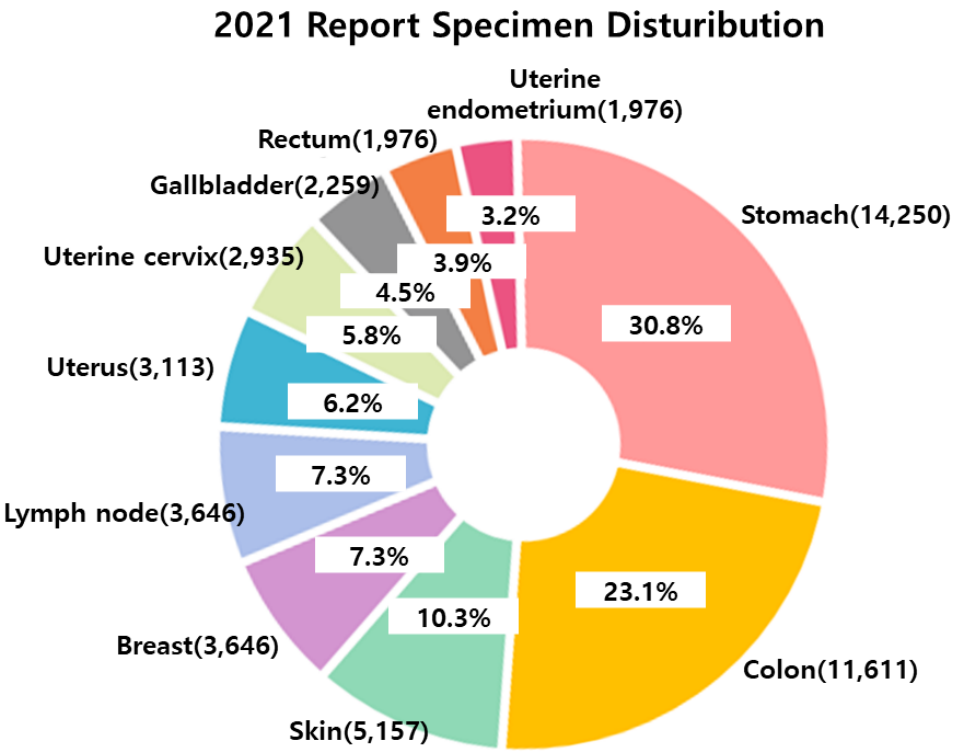

Figure 2. Distribution of Specimens from the Top 10 in the 2021 Pathology Report Dataset.

Zero-shot performance showed that the PLIP model demonstrated a better F1 score (0.33) compared to the CLIP model (0.06). PLIP identified the Fallopian tube (0.53), which CLIP (<0.01) failed to do (Table 1).

Linear probing results indicated that PLIP achieved a higher F1 score (0.81) compared to CLIP (0.54). Both models failed to identify the lymph node (<0.01) and rectum (<0.01). The duodenum was excluded from the test set due to its absence (Table 1).

Table 1. Performances of CLIP and PLIP.

	Zero-shot			Linear probing		
	Num of images	CLIP	PLIP	Num of images	CLIP	PLIP
		F1-score			F1-score	
Breast	147	0.03	0.30	44	0.20	0.88
Colon	254	0.09	0.09	84	0.50	0.81
Duodenum	6	0.02	0.00	-	-	-
Fallopian tube	58	**0.00**	**0.53**	15	0.75	0.97
Lung	24	0.25	0.00	5	0.00	0.56
Lymph node	25	0.08	0.00	8	**0.00**	**0.00**
Rectum	34	0.00	0.18	10	**0.00**	**0.00**
Skin	44	0.02	0.29	10	0.00	0.84
Stomach	321	0.07	0.63	95	0.59	0.83
Thyroid gland	127	0.04	0.03	43	0.96	0.97
Uterine cervix	18	0.00	0.09	6	0.00	0.29
Uterine endometrium	20	0.05	0.00	4	0.08	0.53
Total	1,078	**0.06**	**0.33**	324	**0.54**	**0.81**

Conclusion

This study aimed to apply pathology data from AUSOM to both the CLIP and PLIP models, evaluating and comparing their performance in predicting the organ source of the tissue samples based solely on pathology images.

In conclusion, this study highlights the value of training models with pathology images by demonstrating the superior performance of the PLIP model. Future research should focus on overcoming the identified limitations by incorporating more diverse tissue types and developing techniques to handle high-resolution WSIs, ultimately improving the accuracy and utility of pathology foundation models.

Acknowledgment

This research was funded by a grant from the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea (grant number: HR16C0001) and this research was supported by a Government-wide R&D Fund project for infectious disease research (GFID), Republic of Korea (grant number: HG22C0024, KH124685).

References

1. Madabhushi A. Digital pathology image analysis: opportunities and challenges. Imaging Med. 2009;1(1):7-10.

2. Amol Singh, Robert S. Ohgami, Super-Resolution Digital Pathology Image Processing of Bone Marrow Aspirate and Cytology Smears and Tissue Sections, Journal of Pathology Informatics, Volume 9, Issue 1, 2018, 48.
3. Huang, Z., Bianchi, F., Yuksekgonul, M. et al. A visual–language foundation model for pathology image analysis using medical Twitter. Nat Med 29, 2307–2316 (2023).
4. Radford, Alec, et al. "Learning transferable visual models from natural language supervision." International conference on machine learning. PMLR, 2021.