Exploring the Potential of Histopathology Foundation Models for Slide-Level Classification

Ali Balapour (12458964)
ali20004@student.ubc.ca

Abstract

Computational Pathology (CPath) is a rapidly evolving field in medical image analysis. The growth in machine learning, deep learning, and computer vision has empowered computer scientists to develop tools that aid medical experts in analyzing pathology images for diagnosis, detection, and treatment suggestions. Currently, foundation models in deep learning, particularly in computational pathology, are being explored for their multi-modal capabilities. This work aims to investigate and enhance the performance of the slide-level classification by using three novel foundation model introduced recently. Evaluation of the model will be performed using established metrics like F1 score, AUC-ROC, and Cohen Kappa score. Despite challenges such as the need for large and diverse datasets and substantial computational resources, this work seeks to uncover capabilities and effectiveness of the foundation models in slide-level classification as a core task in computational pathology.

1 Introduction

Analyzing pathology images is crucial for diagnosing cancer, predicting survival rates, counting cells, and segmenting tumors. However, the large size of whole slide images (WSI) in pathology poses a significant challenge for manual analysis. Computational pathology (CPath) can streamline this process by using computational approaches to analyze and model pathology images [17]. Recent advancements in Artificial Intelligence (AI) and Machine Learning (ML) are being gradually incorporated into computational pathology. This includes the concept of self-supervised contrastive learning, first introduced by [9], and later applied to CPath [10]. One of the emerging trends in AI is the use of foundation models. After the introduction of models like CLIP [26] and GPT-3 [6], these types of models have been adopted in CPath with the development of models such as the Pathology Language–Image Pretraining (PLIP) [18].

Foundation models are the most significant set of models in the current state of machine learning research. This family of models is trained on a broad range of data to perform a wide array of downstream tasks. Building foundation models is costly and requires intensive computational resources. Additionally, these models require a vast amount of data from different modalities. In contrast, adapting and inferring foundation models for specific downstream tasks are less expensive and generally lead to more accurate models. The main idea of foundation models is to use them as a starting point to develop machine learning models that can be applied to new applications more effectively and quickly, instead of training from scratch.

In CPath, a number of novel foundation models are introduced in recent years. These models are typically trained on large volumes of data, particularly in the image modality. Some are also trained on pairs of image and text modalities. Having been exposed to millions or even billions of images during their pre-training stage, these models have a deep understanding of images. This makes them useful in CPath tasks that require image encoders to convert pathology images into embeddings. One of this tasks is slide-level classification by using Multi-Instance Learning (MIL) [16]. Traditional MIL models often employ image encoders pre-trained on the ImageNet dataset [12], which aren't particularly effective with pathology images. However, using foundation models as image encoders can enhance the performance of MIL models in slide-level classification. This work will employ three different novel CPath foundation models: PLIP [18], CONCH [21], and UNI [8], as encoders for the state-of-the-art TransMIL [31] method. This will be used for slide-level classification on the Prostate cANcer graDe Assessment (PANDA) [7] dataset.

To summarize, the contributions of this work will be:

- Applying foundation models to a well-known dataset for slide-level classification and comparing the results with existing methods using Cohen's Kappa, AUC-ROC, and F1 score.
- Utilizing explainability techniques to highlight the parts of the image that different encoders focus on and examining the effect of that part of the image on the final result.
- Exploring the challenges, limitations, and potential future directions of using foundation models in core CPath tasks.

2 Related Works

Foundation models are general-purpose, large ML models trained on broad data, enabling their application across a wide variety of tasks [4]. The first instances of foundation models were introduced in Natural Language Processing with BERT [29] and GPT-n [27] models. Besides text data, foundation models have explored other data modalities such as image [28][30][26], music [11], robotic control [5], and mathematics [3]. The potential of foundation models can be useful in medical-related fields such as medical image analysis.

Zhang et al. [36] discussed the potential and challenges of foundation models in medical image analysis and provided an illustration of the spectrum of foundation models in this field. Moor et al. [23] proposed Generalist Medical AI (GMAI), a new paradigm in medical AI based on foundation models. GMAIs are foundation models capable of performing different tasks with limited supervision or access to labeled data. Azad et al. [2] discussed foundation models in medical image analysis and the potential future of these models in this research field.

Huang et al. [18] introduced PLIP, a multi-modal foundation for pathology images by first collecting 208,414 pairs of pathology images (OpenaPath dataset) and texts from Twitter, then pre-training the CLIP [26] model on the collected set. PLIP performs exceptionally well on classification in zero-shot and supervised on 4 external datasets. As PLIP is multi-modal on text and image, it can be used for retrieval based on image or natural language search. The authors provided a demo that can be used for image retrieval and can be used to educate pathologists by finding similar cases. As mentioned in the paper, PLIP can potentially be used in other tasks, so in this study, we will investigate PLIP's abilities in other tasks and datasets.

RUDOLFV [13] and Virchow [34] are innovative foundation models that achieve state-of-the-art performance in computational pathology tasks. Both models utilize the DINOv2 self-supervised learning method. Virchow is pre-trained on 1.5 million slides, while RUDOLFV is pre-trained on 750 million patches. However, these two models only use image modality data. In contrast, PLIP utilizes both image and text data for pre-training. PLIP's backbone is ViT-B-32 [14], while RUDOLFV and Virchow use ViT-H.

UNI [8], a general-purpose, self-supervised vision encoder for pathology, is pretrained on Mass-100K, a dataset of over 100 million tissue patches from various sources. The pretraining uses a self-supervised learning approach, DINOv2 [24], which provides strong representations for downstream tasks without further fine-tuning. UNI is assessed on 34 clinical tasks across anatomical pathology, including cancer detection, grading, and subtyping, biomarker screening, organ transplant assessment, and pan-cancer classification tasks.

CONCH [21] is a visual-language foundation model developed for histopathology images and biomedical text. It uses an image encoder, a text encoder, and a multimodal fusion decoder, trained through contrastive alignment objectives and a captioning objective. The model's capabilities have been tested on various tasks, including image classification, cross-modal retrieval, image segmentation, and image captioning, using 14 diverse benchmarks.

3 Method

3.1 Multi-instance Learning

Multi-instance learning (MIL) is a form of supervised learning, useful for problems where individual labels for training instances are hard to obtain, but group labels are available. MIL can be applied to problems such as slide-level classification in computational pathology, where whole slide images have a general label and patches do not. Among different MIL methods proposed in recent years, TransMIL [31] has shown significant results.

Most current MIL methods are based on the independent and identical distribution hypothesis, neglecting the correlation among different instances. Correlated MIL is proposed in the TransMIL framework to address this issue. By using the self-attention mechanism, conditional positional encoding, local information fusion, and deep feature aggregation modules, morphological and spatial information are captured and used to classify tissue images at the whole slide level.

3.2 Problem Formulation

We have whole slide images (WSIs) as bags, and each of these bags has corresponding instances which are patch images of the tissue in the slide. A weak label will be assigned to each patch, and then based on these weak labels, the general label for each bag will be generated. In general, the MIL is trying to aggregate the information of each instance to decide for the bag.

Consider we have b bags (slides) and slide X_i has n instances (patches): $\{x_{i,1}, x_{i,2}, \ldots, x_{i,n}\}$. The instance-level labels $\{y_{i,1}, y_{i,2}, \ldots, y_{i,n}\}$ are unknown, and we just know bag-level label which is Y_i, for $i = 1, \ldots, b$. By assuming we have binary MIL classification we can define:

$$
\begin{aligned}
&Y_i = \left\{ \begin{array}{l} 0, \text{ iff } \sum y_{i,j} = 0 \quad y_{i,j} \in \{0,1\}, j = 1 \ldots n \\ 1, \text{ otherwise} \end{array} \right. \\
&\hat{Y}_i = S\left(\mathbf{X}_i\right)
\end{aligned}
$$

where S is scoring function and $\hat{Y}_i$ is the prediction. The goal of MIL is finding the S function to predict the label for the slide based on the patches. [31]

3.3 TransMIL

In the TransMIL paper, the authors' contribution was the proposal of a module known as TPT (Transformer-Positional encoder-Transformer), which includes two transformer layers and a position encoding layer. The role of the transformer layers is to aggregate morphological information, such as quantitative representations of the shape, structure, and visual patterns of cells and tissue structures. The positional encoding layer is designed to capture positional and relational understanding about the arrangement and distribution of cells and tissue structures in the image.

In TransMIL, initially, captured patches of a slide are given to a customized ResNet50 pre-trained on the ImageNet dataset [link], and the representation of each patch is extracted as a d-dimensional token. Then we have the Squaring step, in which some of the first tokens in the sequence of patches are selected and placed at the end of the sequence. This squaring is for the sake of the TPT module. After this step, the output is given to the TPT module which, in the first part, applies a transformer layer to the input. This transformer layer uses Nystrom attention, which is proposed in [35]. Using the Nyströmformer reduces the complexity of calculating attention to the order of O(N). This change makes the architecture suitable for classifying slides, as the number of patches might be large.

After the first layer of the Transformer, there is a Pyramid Position Encoding Generator (PPEG) module. This module is designed to add positional features to the representation of the patches, allowing the model to exploit the relative and positional information of the patches. In this module, the patch tokens are converted into a 2D array. Then, different sized convolution layers are applied to this 2D array, and the output is flattened to create a sequence. After the PPEG module, there is another transformer layer identical to the first. At the end, by using the class token, we can employ a multi-layer perceptron (MLP) to generate the final label of the WSI.

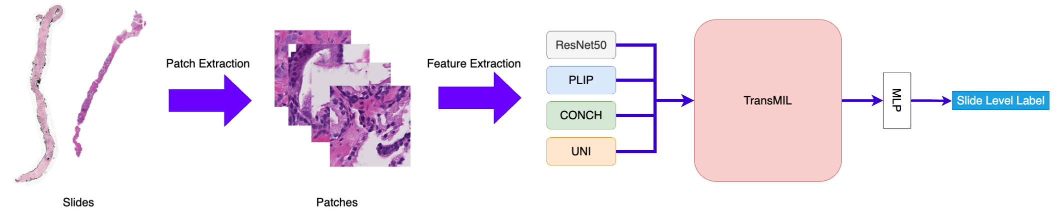

Figure 1: Schematic of the proposed pipeline for slide-level classification

3.4 Image Encoder

In the first step of TransMIL, patch images are given to a ResNet50 feature encoder, which was pre-trained on the ImageNet dataset. ImageNet is a large dataset consisting of natural images and is the most well-known and widely used dataset in computer vision for the task of image classification. Using models pre-trained on this dataset is a best practice in various core tasks of computer vision. A large number of proposed methods in CPath use ImageNet pre-trained models in their pipeline. However, one major issue with the models pre-trained on this dataset is the difference between natural images and pathology images. These differences are in shape, resolution, color, and dimensionality. These fundamental differences might lead to the ineffectiveness of ImageNet pre-trained models in CPath applications [1].

To address this challenge, in this work, we are attempting to use a number of well-known foundational models in histopathology to extract embeddings to give to TransMIL. These foundational models are PLIP, UNI, and CONCH. These models have shown excellent performance when used as feature encoders for different applications such as zero-shot patch-level classification.

3.5 Pipeline

To perform slide-level classification using the MIL method, we first need to prepare bags and instances. Each WSI is considered a bag, and each patch in the slide is considered an instance of the bag. We used the CLAM repository[1] [22] to extract 256*256 patches from each WSI. This tool first segments the entire tissue within the WSI, then tiles the segmented part and saves the coordinates of each patch. Saving coordinates instead of each patch results in less storage and memory usage. In the next step, we used CLAM to extract features using ResNet50 (pre-trained on ImageNet), PLIP, CONCH, and UNI image encoders. All encoders were provided by CLAM, except for PLIP, for which we customized the tool to provide support. After generating embeddings, they are given to the TransMIL model for processing. The general structure of our pipeline is depicted in Figure 1.

4 Experiments and Results

4.1 Prostate cANcer graDe Assessment (PANDA) Dataset

The Prostate Cancer Grade Assessment (PANDA) [7] dataset is a groundbreaking resource for developing and evaluating automated deep learning systems for prostate cancer diagnosis and grading. Prostate cancer is the second most common cancer among males worldwide, with over 1 million new diagnoses reported every year. Accurate diagnosis and grading of prostate cancer is crucial, as it impacts treatment decisions and patient outcomes. However, the current manual grading process by medical experts suffers from significant inter and intra-observer variability. The PANDA dataset is a large public whole-slide image dataset available for prostate cancer grading, consisting of around 10,600 H&E-stained biopsy slides from two medical centres.

Gleason grading is utilized to determine the stage of prostate cancer, providing vital prognostic information. Uropathologists classify tumors into different Gleason growth patterns based on the histological architecture of the tumor tissue. Based on the distribution of these Gleason patterns, biopsy specimens are categorized into one of five groups. These groups are commonly referred to as International Society of Urological Pathology (ISUP) grade groups [33][15]. Gleason score and

[1] https://github.com/mahmoodlab/CLAM

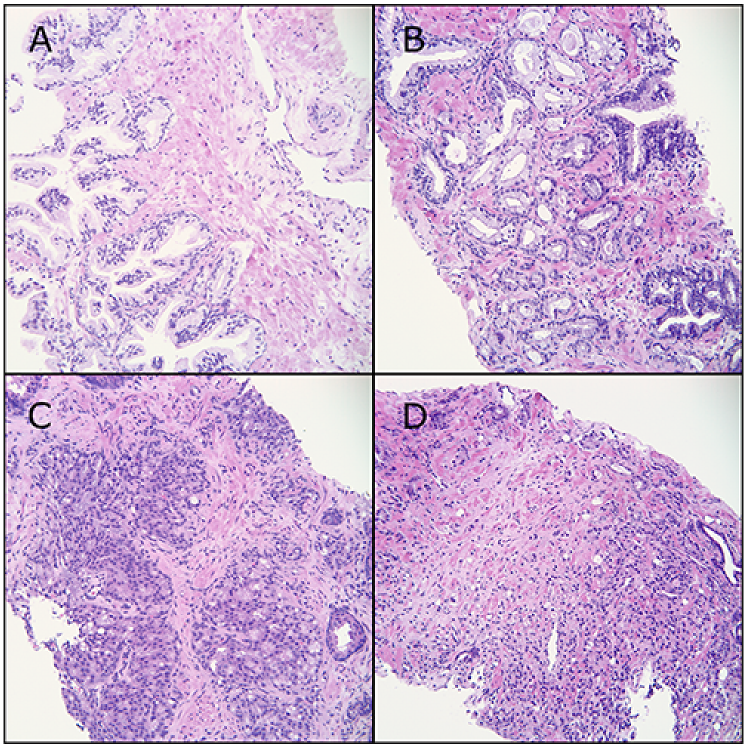

Figure 2: Examples of patterns used for grading. (A) Healthy sample. (B) Gleason pattern 3, (C) Gleason pattern 4, and (D) Gleason pattern 5. [7]

ISUP grades are provided for each slide in the PANDA dataset. The ISUP grade will be used as the label of each slide for training. Figure 2 depicts 4 sample patch of PANDA dataset with different ISUP grades.

The PANDA dataset includes mask annotations for each slide, highlighting different types of tissue. One approach for cancer prediction using this dataset involves applying segmentation models, aggregating the results, and using them to predict a score for each slide. However, in this study, we aim to use the given ISUP grade for each slide and apply a Multiple Instance Learning (MIL) method to predict this score without relying on mask annotations. This task presents more of a challenge as we only have a single label instead of complete tissue masks.

4.2 Experiment Design

We randomly selected 10% of the PANDA dataset as a test set and performed 3-fold cross-validation on the remaining part to determine how well the histopathology foundation models performed in the task of slide-level classification. For all experiments, we used the Kaggle environment with one Nvidia P100 GPU. We utilized the PyTorch library [25], along with Pandas, NumPy, and Scikit-Learn for implementing and evaluating our method. For extracting patches and generating image embeddings with foundation models, we customized the CLAM repository, and for training the model, we modified the original TransMIL repository. Guidelines on how to run the model are available through the provided GitHub repository[2]. To compare the results of each configuration, we used Accuracy, Cohen Kappa score, F1 score, Recall, Precision, Specificity, and AUC-ROC. The results are in Table 1.

4.3 Results

According to Figure 3, when comparing different encoders using F1, Cohen Kappa, and AUC-ROC scores, UNI and CONCH encoders outperform PLIP and ResNet. UNI performs slightly better

[2] https://github.com/alibalapour/Foundation-TransMIL

Table 1: Results of using different encoders in TransMIL on test set with using 3-fold cross validation.

Model	Fold	Accuracy	Cohen Kappa	F1 Score	Recall	Precision	Specificity	AUC-ROC
ResNet	0	0.55833334	0.47000003	0.5367328	0.55833334	0.54844666	0.9116667	0.86833835
	1	0.5625	0.47499996	0.54713887	0.5625	0.5695398	0.9125	0.8731767
	2	0.5325	0.439	0.49818212	0.5325	0.5312007	0.9065	0.86481
PLIP	0	0.5391667	0.44700003	0.52014613	0.5391667	0.5305311	0.90783334	0.8702167
	1	0.5466667	0.45599997	0.5252101	0.5466666	0.533704	0.90933335	0.87239504
	2	0.5525	0.463	0.5346317	0.5525	0.54921925	0.9105	0.86925256
CONCH	0	0.6041667	0.52500004	0.58413243	0.6041667	0.59760225	0.92083335	0.8945259
	1	0.62583333	0.551	0.6126064	0.62583333	0.62721395	0.92516667	0.9037384
	2	0.5975	0.51699996	0.5711872	0.59749997	0.6054673	0.9195	0.8974775
UNI	0	0.6433333	0.572	0.62920666	0.6433333	0.6341343	0.9286667	0.91348666
	1	0.62916666	0.555	0.60921824	0.62916666	0.6454349	0.92583334	0.9126475
	2	0.6625	0.595	0.6557206	0.66249996	0.6550435	0.93249995	0.92115253

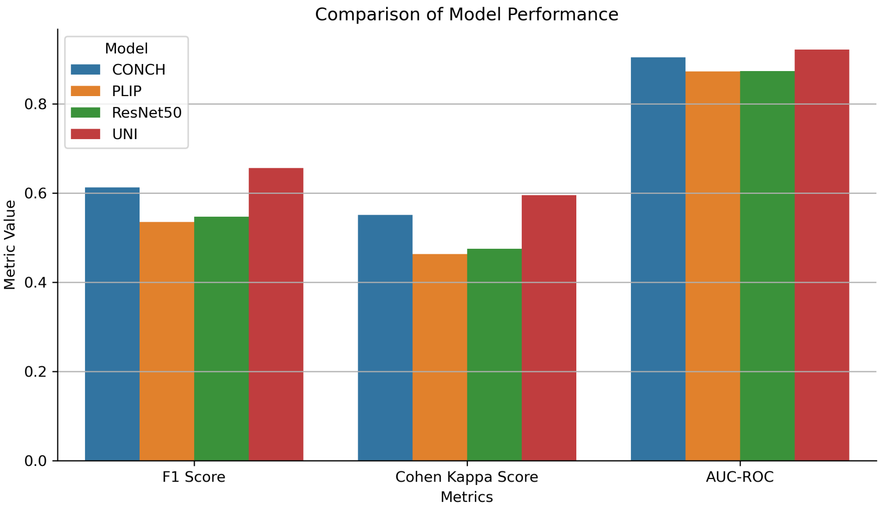

Figure 3: Results of best found model in 3-fold cross validation on the test set based on F1, Cohen Kappa, and AUC-ROC score.

than CONCH, while ResNet and PLIP perform similarly based on these metrics. As per Table 2, UNI shows an improvement of approximately 11%, 10%, and 5% in Cohen Kappa, F1 score, and AUC-ROC, respectively, compared to the ResNet encoder. CONCH exhibits an improvement of 7% in Cohen Kappa, 6% in F1 score, and 3% in AUC-ROC over ResNet. However, there is no significant improvement in the PLIP model compared to ResNet.

Our results are significantly lower compared to state-of-the-art works such as [20] and [32]. In [32], the authors employed a similar strategy using a Multiple Instance Learning (MIL) method and a custom contrastive learning-based image encoder. They achieved 84.9% accuracy and a 78.6% Cohen Kappa score in their 5-fold cross-validation. Compared to our work, they extracted 512*512 patches and used a different MIL method, which might explain why their model outperformed ours. However, our model could improve its performance with more hyper-parameter tuning on TransMIL or by customizing TransMIL to increase its complexity. Conversely, [20] used ground truth masks of slides and achieved 85.3% accuracy on a portion of the PANDA dataset through segmentation. Given the additional ground truth information used, it's understandable that this approach yielded better results.

Table 2: Summary of results of using different encoders in TransMIL on test set.

Model	Cohen Kappa	F1 Score	AUC-ROC
ResNet	0.461 (0.439-0.474)	0.527 (0.498-0.547)	0.868 (0.864-0.873)
PLIP	0.455 (0.447-0.463)	0.526 (0.52-0.534)	0.87 (0.869-0.872)
CONCH	0.531 (0.516-0.551)	0.589 (0.571-0.629)	0.898 (0.894-0.903)
UNI	0.574 (0.555-0.595)	0.631 (0.609-0.655)	0.915 (0.912-0.921)

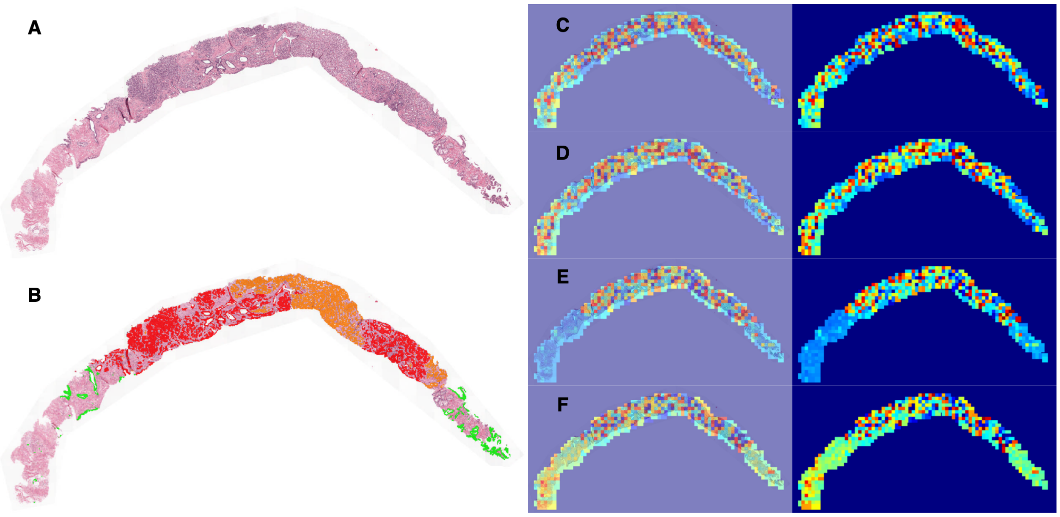

Figure 4: Comparing the original slide (A) and original mask (B) with the attention maps of ResNet50 (C), PLIP (D), CONCH (E), and UNI (F) obtained with TransMIL. In the original mask, red and orange represent cancerous regions. While ResNet50 and PLIP outputs offer little distinction between cancerous (center of tissue) and non-cancerous regions (ends of tissue), CONCH and UNI can more effectively distinguish these regions.

4.4 Visualizations

To demonstrate the effect of using different image encoders on the result of the TransMIL model, we present the attention map of each encoder on a slide sample. This is achieved by using attention values for each patch and plotting it on the thumbnail of the slide sample. In Figure 4, different maps for each encoder can be seen, compared to the original mask of the sample. In the provided sample, the tumor is primarily in the center of the tissue, and the amount of tumors is less in the tails (especially the left tail). As evident in the attention maps, the ResNet and PLIP encoders struggle to differentiate effectively between tumor and non-tumor regions. However, with the CONCH and UNI encoders, the tails are less attended compared to the center, and there is a clear distinction between these two regions. This could explain why the CONCH and UNI encoders perform better, as we discovered in the results section.

In Figure 5, to assess the encoders' ability to distinguish the extracted features of each class, we selected five sample slides per class. We then applied t-SNE to the features of the patches from the chosen slides and visualized them on a 2D scatter plot. The features extracted from the UNI model exhibited the best differentiation between classes. It's worth noting that within the cancerous class slides, some parts may be normal and not related to the tumor. Consequently, the sample patches from different classes displayed on the scatter plots might appear similar and cluster together.

5 Challenges, Limitations, and Future Directions

Given the large size of histopathology slides, processing these images poses a challenge and requires sufficient computational resources. However, the existence of repositories like CLAM has helped us overcome this challenge, as it offers efficient methods for processing slides on a conventional system.

One limitation of the foundation models available is they are trained on millions of data from different organs and cancer type. Tissue of each cancer type and organ could be different and this might

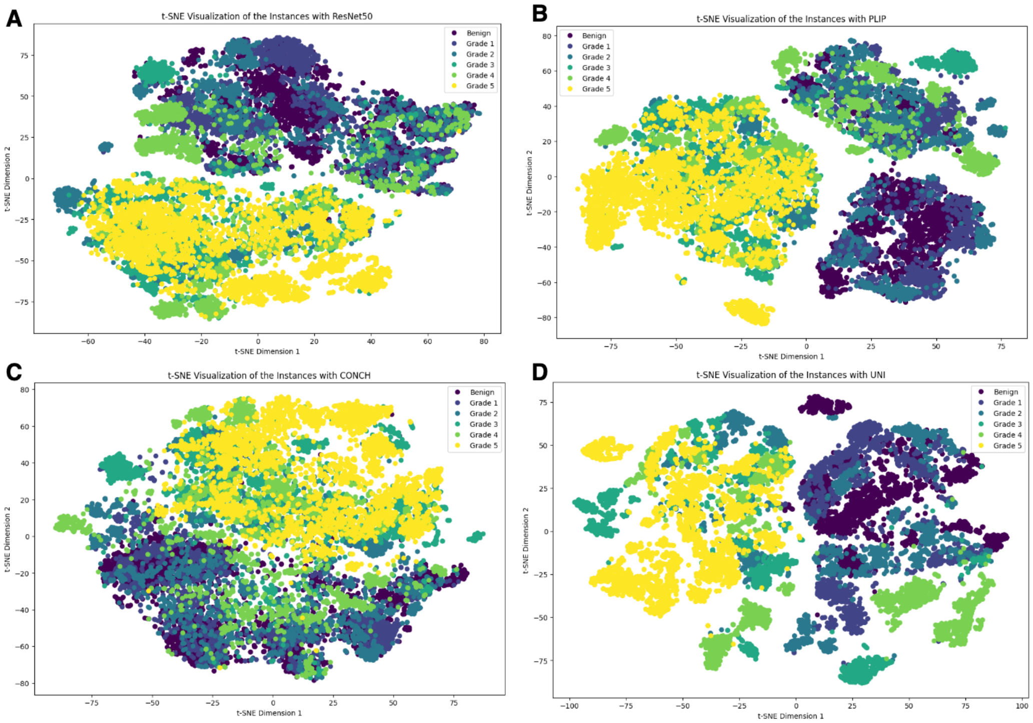

Figure 5: Visualization of features extracted from different encoders by using t-SNE. (A) represents features extracted from ResNet50, (B) features from PLIP, (C) features from CONCH, and (D) features from UNI.

lead the model to not perform good on data from different organ because of domain shift. So, one solution might be considering this domain shift while pre-training the foundation model or making the model more capable of understanding the different tissue types. Another solution is pre-training the foundation model with similar data before applying it in downstream task. One limitation of our work is that the PANDA dataset consists of slides from two different centres, resulting in a domain shift between these two types of data. This domain shift should be considered in our work. However, for simplicity, we did not take this factor into account. Another constraint of our study is the limited access to the internal test set of the PANDA dataset, as the PANDA challenge host restricts its accessibility. If we could evaluate our models using this internal test set, it would enable a more effective comparison of our results with those from other studies.

Regarding future directions, we can consider using other notable CPath datasets to assess the performance of foundation models. Additionally, pre-training the foundation model on data similar to the target dataset of the downstream task could be a worthwhile step to investigate. Other MIL methods may also be valuable for future work. An exciting possibility could be to combine different foundation models, leveraging their unique strengths while mitigating their weaknesses. Finally, it would be beneficial to explore whether these models could be applied to other image-based tasks within pathology, such as segmentation, image synthesis, and image captioning. As an example, for segmentation, we can utilize the Segment Anything (SAM) [19] architecture, and employ the image encoders such as UNI or CONCH. We will then use PLIP or any vision language histopathology foundation model as the text encoder of the modified SAM model.

6 Conclusion

In this work, we aimed to explore the ability of histopathology foundation models as image encoders in slide-level classification tasks. We showed that by using these models instead of conventional methods such as ResNet50 pre-trained on ImageNet, the performance of the state-of-the-art methods in this task can significantly improve. Using UNI and CONCH foundation models as image encoders in the TransMIL model can improve the Kohen Kappa score by around 11% and 7% respectively, compared to ResNet50. On the other hand, not all foundation models may be suitable for this task. Our experiments showed that the PLIP foundation model does not have superiority over ResNet50. Our findings indicate that the choice of foundation model can significantly impact the final results and that selecting the appropriate model based on the characteristics of the task and dataset can lead to better outcomes.

References

[1] Laith Alzubaidi, Muthana Al-Amidie, Ahmed Al-Asadi, Amjad J. Humaidi, Omran Al-Shamma, Mohammed Abdulraheem Fadhel, Jinglan Zhang, Jesus Santamaría, Ye Duan, Keyvan Farahani, and Bardia Yousefi. Novel transfer learning approach for medical imaging with limited labeled data. *Cancers*, 13, 2021.

[2] Bobby Azad, Reza Azad, Sania Eskandari, Afshin Bozorgpour, Amirhossein Kazerouni, Islem Rekik, and Dorit Merhof. Foundational models in medical imaging: A comprehensive survey and future vision. *ArXiv*, abs/2310.18689, 2023.

[3] Zhangir Azerbayev, Hailey Schoelkopf, Keiran Paster, Marco Dos Santos, Stephen McAleer, Albert Q. Jiang, Jia Deng, Stella Biderman, and Sean Welleck. Llemma: An open language model for mathematics, 2023.

[4] Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie Chen, Kathleen Creel, Jared Quincy Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Ben Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, Julian Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Rob Reich, Hongyu Ren, Frieda Rong, Yusuf Roohani, Camilo Ruiz, Jack Ryan, Christopher Ré, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishnan Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. On the opportunities and risks of foundation models, 2022.

[5] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, Pete Florence, Chuyuan Fu, Montse Gonzalez Arenas, Keerthana Gopalakrishnan, Kehang Han, Karol Hausman, Alexander Herzog, Jasmine Hsu, Brian Ichter, Alex Irpan, Nikhil Joshi, Ryan Julian, Dmitry Kalashnikov, Yuheng Kuang, Isabel Leal, Lisa Lee, Tsang-Wei Edward Lee, Sergey Levine, Yao Lu, Henryk Michalewski, Igor Mordatch, Karl Pertsch, Kanishka Rao, Krista Reymann, Michael Ryoo, Grecia Salazar, Pannag Sanketi, Pierre Sermanet, Jaspiar Singh, Anikait Singh, Radu Soricut, Huong Tran, Vincent Vanhoucke, Quan Vuong, Ayzaan Wahid, Stefan Welker, Paul Wohlhart, Jialin Wu, Fei Xia, Ted Xiao, Peng Xu, Sichun Xu, Tianhe Yu, and Brianna Zitkovich. Rt-2: Vision-language-action models transfer web knowledge to robotic control, 2023.

[6] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020.

[7] Wouter Bulten, Kimmo Kartasalo, Po-Hsuan Cameron Chen, Peter Ström, Hans Pinckaers, Kunal Nagpal, Yuannan Cai, David F. Steiner, Hester H. van Boven, Robert Vink, Christina A. Hulsbergen–van de Kaa, Jeroen A. van der Laak, Mahul B. Amin, Andrew J. Evans, Theodorus H van der Kwast, Robert Allan, Peter A. Humphrey, Henrik Grönberg, Hemamali Samaratunga, Brett Delahunt, Toyonori Tsuzuki, Tomi Häkkinen, Lars Egevad, Maggie Demkin, Sohier Dane, Fraser Tan, Masi Valkonen, Greg S Corrado, Lily H. Peng, Craig H. Mermel, Pekka Ruusuvuori, Geert J. S. Litjens, Martin Eklund, Américo Aslı Xavier Katerina Vincent Guilherme Paromita Gü Brilhante Çakır Farré Geronatsiou Molinié Pereira, Américo Delgado Brilhante, Aslı Çakır, Xavier Farré, Katerina Geronatsiou, Vincent Molinie, Guilherme Pereira, Paromita Roy, Günter Saile, Paulo Guilherme de Oliveira Salles, Ewout Schaafsma, Joëlle Tschui, Jorge Billoch-Lima, Emíio M. Pereira, M. Zhou, Shujun He, Sejun Song, Qing Sun, Hiroshi Yoshihara, Taiki Yamaguchi, Kosaku Ono, Tao Shen, Jianyi Ji, Arnaud Roussel, Kairong Zhou, Tianrui Chai, Nina Weng, Dmitry A. Grechka, Maxim V. Shugaev, Raphael Kiminya, Vassili A. Kovalev, Dmitry Voynov, Valery Malyshev, Elizabeth Lapo, Manolo Quispe Campos, Noriaki Ota, Shinsuke Yamaoka, Yusuke Fujimoto, Kentaro Yoshioka, Joni Juvonen, Mikko Tukiainen, Antti Karlsson, Rui Guo, Chia-Lun Hsieh, I. S. Zubarev, Habib S. T. Bukhar, Wenyuan Li, Jiayun Li, W. Speier, Corey W. Arnold, Kyungdoc Kim, Byeonguk Bae, Yeong Won Kim, Hong-Seok Lee, and Jeonghyuk Park. Artificial intelligence for diagnosis and gleason grading of prostate cancer: the panda challenge. *Nature Medicine*, 28:154 – 163, 2022.

[8] Richard J Chen, Tong Ding, Ming Y. Lu, Drew F. K. Williamson, Guillaume Jaume, Andrew H. Song, Bowen Chen, Andrew Zhang, Daniel Shao, Muhammad Shaban, Mane Williams, Lukas Oldenburg, Luca L Weishaupt, Judy J. Wang, Anurag Vaidya, Long Phi Le, Georg Gerber, Sharifa Sahai, Walt Williams, and Faisal Mahmood. Towards a general-purpose foundation model for computational pathology. *Nature medicine*, 2024.

[9] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations, 2020.

[10] Ozan Ciga, Tony Xu, and Anne L. Martel. Self supervised contrastive learning for digital histopathology, 2021.

[11] Jade Copet, Felix Kreuk, Itai Gat, Tal Remez, David Kant, Gabriel Synnaeve, Yossi Adi, and Alexandre Défossez. Simple and controllable music generation, 2024.

[12] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, K. Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.

[13] Jonas Dippel, Barbara Feulner, Tobias Winterhoff, Simon Schallenberg, Gabriel Dernbach, Andreas Kunft, Stephan Tietz, Philipp Jurmeister, David Horst, Lukas Ruff, Klaus-Robert Müller, Frederick Klauschen, and Maximilian Alber. Rudolfv: A foundation model by pathologists for pathologists, 2024.

[14] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021.

[15] Jonathan I. Epstein, Lars Egevad, Mahul B. Amin, Brett Delahunt, John R. Srigley, and Peter A. Humphrey. The 2014 international society of urological pathology (isup) consensus conference on gleason grading of prostatic carcinoma: Definition of grading patterns and proposal for a new grading system. *The American Journal of Surgical Pathology*, 40:244–252, 2015.

[16] Michael Gadermayr and Maximilian Ernst Tschuchnig. Multiple instance learning for digital pathology: A review on the state-of-the-art, limitations & future potential. *Computerized medical imaging and graphics : the official journal of the Computerized Medical Imaging Society*, 112:102337, 2022.

[17] Mahdi S. Hosseini, Babak Ehteshami Bejnordi, Vincent Quoc-Huy Trinh, Danial Hasan, Xingwen Li, Taehyo Kim, Haochen Zhang, Theodore Wu, Kajanan Chinniah, Sina Maghsoudlou, Ryan Zhang, Stephen Yang, Jiadai Zhu, Lyndon Chan, Samir Khaki, Andrei Buin, Fatemeh

Chaji, Ala Salehi, Bich Ngoc Nguyen, Dimitris Samaras, and Konstantinos N. Plataniotis. Computational pathology: A survey review and the way forward, 2024.

[18] Zhi Huang, Federico Bianchi, Mert Yuksekgonul, Thomas J. Montine, and James Y Zou. A visual–language foundation model for pathology image analysis using medical twitter. *Nature Medicine*, 29:2307–2316, 2023.

[19] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross B. Girshick. Segment anything. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3992–4003, 2023.

[20] Ma Liang, Chen Hao, and Gong Ming. Prostate cancer grade using self-supervised learning and novel feature aggregator based on weakly-labeled gbit-pixel pathology images. *Appl. Intell.*, 54:871–885, 2023.

[21] Ming Y Lu, Bowen Chen, Drew FK Williamson, Richard J Chen, Ivy Liang, Tong Ding, Guillaume Jaume, Igor Odintsov, Long Phi Le, Georg Gerber, et al. A visual-language foundation model for computational pathology. *Nature Medicine*, 30:863–874, 2024.

[22] Ming Y Lu, Drew FK Williamson, Tiffany Y Chen, Richard J Chen, Matteo Barbieri, and Faisal Mahmood. Data-efficient and weakly supervised computational pathology on whole-slide images. *Nature Biomedical Engineering*, 5(6):555–570, 2021.

[23] Michael Moor, Oishi Banerjee, Zahra F H Abad, Harlan M. Krumholz, Jure Leskovec, Eric J. Topol, and Pranav Rajpurkar. Foundation models for generalist medical artificial intelligence. *Nature*, 616:259–265, 2023.

[24] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2024.

[25] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library, 2019.

[26] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021.

[27] Alec Radford and Karthik Narasimhan. Improving language understanding by generative pre-training. 2018.

[28] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation, 2021.

[29] Anna Rogers, Olga Kovaleva, and Anna Rumshisky. A primer in bertology: What we know about how bert works, 2020.

[30] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2022.

[31] Zhuchen Shao, Hao Bian, Yang Chen, Yifeng Wang, Jian Zhang, Xiangyang Ji, and Yongbing Zhang. Transmil: Transformer based correlated multiple instance learning for whole slide image classification, 2021.

[32] Nitin Singhal, Shailesh Soni, Saikiran Bonthu, Nilanjan Chattopadhyay, Pranab Samanta, Uttara Joshi, Amit Jojera, Taher Chhharchhodawala, Ankur Agarwal, Mahesh Desai, and Arvind Prakash Ganpule. A deep learning system for prostate cancer diagnosis and grading in whole slide images of core needle biopsies. *Scientific Reports*, 12, 2022.

[33] Geert J.L.H. van Leenders, Theodorus H. van der Kwast, David J. Grignon, Andrew J. Evans, Glen Kristiansen, Charlotte F. Kweldam, Geert J. S. Litjens, Jesse K. McKenney, Jonathan Melamed, N. Mottet, Gladell P. Paner, Hemamali Samaratunga, Ivo G. Schoots, Jeffry P. Simko, Toyonori Tsuzuki, Murali Varma, Anne Y. Warren, Thomas M. Wheeler, Sean R. Williamson, and Kenneth A. Iczkowski. The 2019 international society of urological pathology (isup) consensus conference on grading of prostatic carcinoma. *The American Journal of Surgical Pathology*, 44:e87 – e99, 2020.

[34] Eugene Vorontsov, Alican Bozkurt, Adam Casson, George Shaikovski, Michal Zelechowski, Siqi Liu, Kristen Severson, Eric Zimmermann, James Hall, Neil Tenenholtz, Nicolo Fusi, Philippe Mathieu, Alexander van Eck, Donghun Lee, Julian Viret, Eric Robert, Yi Kan Wang, Jeremy D. Kunz, Matthew C. H. Lee, Jan Bernhard, Ran A. Godrich, Gerard Oakley, Ewan Millar, Matthew Hanna, Juan Retamero, William A. Moye, Razik Yousfi, Christopher Kanan, David Klimstra, Brandon Rothrock, and Thomas J. Fuchs. Virchow: A million-slide digital pathology foundation model, 2024.

[35] Yunyang Xiong, Zhanpeng Zeng, Rudrasis Chakraborty, Mingxing Tan, Glenn Fung, Yin Li, and Vikas Singh. Nyströmformer: A nyström-based algorithm for approximating self-attention, 2021.

[36] Shaoting Zhang and Dimitris N. Metaxas. On the challenges and perspectives of foundation models for medical image analysis. *Medical image analysis*, 91:102996, 2023.