Testing Foundation Models for Digital Pathology on a Public Prostate Cancer Detection Challenge

Eelke Spijkers
Faculty of Social Sciences
Radboud University
Nijmegen, the Netherlands
eelke.spijkers@ru.nl

Laura Euverman
Faculty of Social Sciences
Radboud University
Nijmegen, the Netherlands
laura.euverman@ru.nl

Mika Klepper
Faculty of Social Sciences
Radboud University
Nijmegen, the Netherlands
mika.klepper@ru.nl

Taha Khaleel
Faculty of Social Sciences
Radboud University
Nijmegen, the Netherlands
taha.khaleel@ru.nl

***Abstract*—The analysis of whole slide images for pathology presents a couple of challenges, mainly associated with time consumption of manual analysis. Furthermore, pathologists often do not agree on a diagnosis. Imaging equipment also tends to present different quality data. Artificial intelligence is expected to be beneficial in overcoming these challenges. The current research aims to show advancements made in model development over the last five years. The main goal is to show that models created specifically for pathological cancer detection purposes are an improvement on the general models used in the PANDAs competition. The current study implements a multiple instance learning based model using Trident, Titan, and Phikon. Three files were used for feature extraction, training, and inference. The final model yielded a quadratic weighted kappa score of 0.76. We did not improve the highest score of the competition from five years ago, 0.94, which is proposed to be due to submission limitations on Kaggle. It is expected that optimized feature extraction methods could increase the score in future research.**

***Index Terms*—Digital Pathology, Prostate Cancer, Cancer Detection, PANDA Challenge, Trident, Titan, Phikon, Multiple instance learning**

I. Introduction

With more and more complex molecular biomarkers for cancer being identified, the process of cancer diagnosis is becoming more intricate. Workflows within oncology have become elaborate, and thus tend to take up more time, even though pathological work often consists of very similar or even repetitive work [1]. As more data on these biomarkers is becoming available, it seems logical to turn to artificial intelligence (AI) models to support pathological workflows. AI models have the potential to increase the uniformity of diagnoses, while saving time spent manually analyzing medical images [1, 2].

Early digital pathology architectures that analyzed images of tissue samples were commonly based on pretrained convolutional neural networks (CNNs), which have been shown to achieve expert level performance [3]. However, training these CNNs requires large datasets which are scarce [4], and commonly ‘out-of-domain’, which causes limitations in the detection capability of details [5]. As such, much recent work has been dedicated to developing better foundation models for pathology, methods for automatic labeling and segmentation of datasets, and self-supervised learning methods [5, 6]. Some developments relevant to the current work include: Multiple-instance learning (MIL), which addresses the scarcity of annotations in whole slide images (WSIs) [4]. Phikon, a transformer-based feature extractor tailored to pathology [4, 5, 7]. Trident, a segmentation pipeline package developed for WSIs, offers scalable batch processing modules that can handle thousands of WSIs from most WSI formats [8]. Titan, a multimodal whole slide representation foundation model developed for digital pathology [9].

We aim to combine the above-mentioned state-of-the-art architectures and models and apply them to a publicly available machine learning challenge from biomedical imaging. Specifically, it aims to do so on the PANDA challenge, originally published on Kaggle in 2020 [10], to test whether these advanced models tailored to digital pathology are able to improve upon the results of earlier models. We test three different variants of consisting of the same basic architecture with different loss functions and data cleaning steps to determine which steps are most effective in our pipeline. We hypothesize that our models will be better able to reliably detect cancer compared to the existing submissions.

II. Methods

A. PANDA Challenge

The aim of the PANDA challenge is to develop a model for prostate cancer detection that predicts the Gleason severity grade of WSIs of tissue samples, as well as the corresponding treatment-guiding International Society of Urological Pathology (ISUP) grade [10]. The Gleason grade is a predictive indicator that is determined by analyzing the microscopic appearance of a tissue sample. The goal is to determine how differentiable the cancerous cells are from the surrounding tissue. A high Gleason grade is commonly indicative of an aggressive form of prostate cancer and thus a worse prognosis [10, 11]. The ISUP grade is a newer standardized metric derived from the Gleason grade and is used to design a treatment plan [12, 13].

The data published for the challenge was gathered by the Radboud University Medical Center (RUMC), Nijmegen, the Netherlands and the Karolinska Institute (KI), Stockholm, Sweden, and included 11,000 whole slide images [10, 12]. The public set included a file with the training and test images, a file with the training labels, and segmentation masks. Earlier submissions noted that the labels are noisy and that images contain artifacts such as pen marks[12]. The noisy labels are attributed in part to the automatic generation of the labels from the RUMC, and the fact that the labels from KI were based on the opinion of a single or few pathologists. Moreover, as the data is coming from separate institutes, there are differences in image size and region segmentation and annotation [12]. The images are large, but also contain much irrelevant patches containing only the empty background. As such, identification and magnification of the patches of interest will be essential.

The PANDA challenge imposed the following constraints on submissions: the model is allowed a maximum of 9 hours run-time when running on a CPU, and a maximum of 6 hours run-time when running on a GPU. The models will run on Kaggle. For the GPU, only the 2x T4 or 1x P100 options are permitted, the use of the v3-8 TPU is only allowed for training models. The submitted model must be able to run without internet access, without using any custom packages. However, it is allowed to use external data that is publicly available, which includes pretrained models [10].

The winning model of the PANDA competition yielded a quadratic weighted kappa score of 0.94085, using p pretrained ResNet model and an ensemble of a 5-fold EfficientNet-B0 and a 5-fold EfficientNet-B1 model [14]. Their main innovation was to remove noisy data by filtering out any training data where the predictions an initial prediction model showed a high disparity between the original and predicted ISUP grade.

Many submissions including the winning solution used feature extraction pipelines composing of (pretrained) CNNs to reduce the entire WSI to feature vectors only patches of interest. This technique proved effective, and is the reason why a pretrained feature extractor is used in the current work.

B. Pipeline

The current study uses a MIL pipeline that includes three steps: slide segmentation, encoding, and finally classification, see Figure 1. The segmentation step uses HEST, a segmentation module that is used to create segmentation masks by distinguishing the background from the foreground, from the Trident package [8]. Feature extraction occurs per patch using Phikon [7] based on the segmentation mask, followed by an aggregation step to slide-level embeddings using Titan [9] to reduce the number of patches and increase computational efficiency. The segmentation and encoding step is used to pre-process the training data for the classification step, and only needs to be run once over the entire training set.

For classification, a Multi-Layer Perceptron (MLP) with 4 layers was used. This model returns a 5-label output. The slide classifier is trained using stratified K-fold cross-validation, and the model that performs best on the validation set is returned. This model is then used to predict the ISUP grade of a slide.

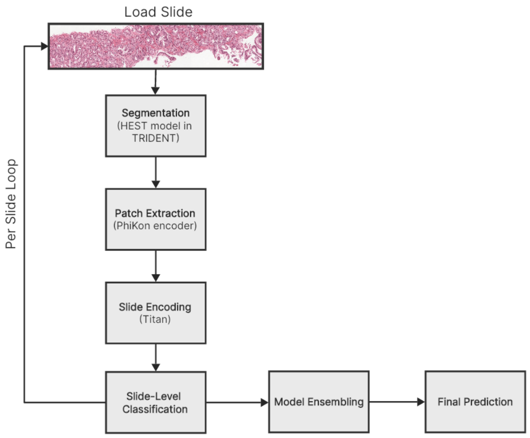

Fig. 1. Inference pipeline for the model

Both CONCH and Phikon were considered as foundational models. As the slides were considerably large, we considered three magnification option for running the models, 5x, 10x, and 20x magnification. The 20x magnification is expected to yield the most detailed results, however, it was soon realized that this option would not be realistic for our current approach due to time constraints. Thus, 10x and 5x magnification were both tested with CONCH and Phikon, with a batch size of 32 for both models. For CONCH we used a patch size of 512 pixels, for which one slide on average took 19.30 seconds when using 10x magnification, and 16-16.30 seconds for 5x magnification. For Phikon we used a patch size of 224 pixels, which yielded an average processing time per slide of 17.30 seconds for 10x magnification, and 15.52 seconds for 5x magnification. We determined all four options possible within the time constraints, when using GPU, with CONCH 10x magnification taking the longest to run on the whole test set, but still was expected to finish within 5.5 hours. This left half an hour for inference, which was considered feasible.

While CONCH seems to outperform Phikon [15], the current study did decide to implement Phikon as foundational model, accompanied by Titan as slide encoder for the aggregation step. The motivation for this mostly stems from the time constraints related to the challenge. When doing a test run on the CONCH model with 10x magnification, the average processing time per slide slowed down, making it not feasible to run within time constraints. We thought higher magnification to be more important to the accuracy than foundation model, which let us to the decision of using Phikon with 10x magnification as foundational model.

C. Models

Three models were designed, with every successive model including one more processing step than the previous, see Table I. The first model consists of a Phikon processing step, followed by a processing step by Titan, and finally an MLP for decision making. This model uses Cross-Entropy loss, with no warm-up or data cleaning. Model two changes the loss function to consider that the labels are ordinal, and also adds warm-up. The last model also includes a label cleaning step where labels that showed too much of a difference between the predicted label and the actual label are removed.

Details	**Model 1**	**Model 2**	**Model 3 (Final Model)**
Architecture	Phikon + Titan with MLP 4 layers	Phikon + Titan with MLP 4 layers	Phikon + Titan with MLP 4 layers
Loss	Cross-Entropy Loss	Binary Cross-Entropy with LogitsLoss with ordinal labels	Binary Cross-Entropy with LogitsLoss with ordinal labels
Scheduler	None	Cosine with warm-up	Cosine with warm-up
Data cleaning	None	None	Label cleaning

TABLE I
MODELS

D. Evaluation

For each image, the model predicts the ISUP grade. The quadratic-weighted kappa is then used to evaluate these predictions, which measures the agreement between the predicted and original ISUP grade. The quadratic-weighted kappa takes into account that the grade is an ordinal value, and thus punished a larger disagreement with a higher ISUP grade more strongly.

III. RESULTS

We have tested three different models, in which each is improved upon the prior. The architecture for all three models is Phikon with Titan on four layers multi layer perceptron. We extended the second model with a scheduler. For the third model we included label cleaning. Table I explains the model structures in more detail.

Models	**Score on Private data**
Model 1	-0.002
Model 2	0.74
Model 3 (Final model)	0.76

TABLE II
RESULTS

IV. DISCUSSION

First implementation of our model, without optimization, yielded a score of -0.002 on the private test set. For this model we used cross entropy and multi-classification. The low score can be explained due to multi-classification not being applicable on ISUP grading. As a high grade should always be considered worse than a low grade, ordinal labeling made more sense for this challenge. The effect was immediately shown in the second model submission. For the second model it was also decided to implement BCE with logit loss instead of cross entropy. This model yielded a submission score of 0.74 on the private dataset. In the third model we implemented label denoising, to which we did see a strong increase in cross validation. Unfortunately this increase was not shown for the private test set, yielding a score of 0.76.

Due to the test data being associated with the Kaggle challenge, we did not have full access to the hidden test set. This did force us to take into account the limitations set for the competition, which also limited our options for optimization. We expect the CONCH model to perform better than the Phikon model, however, this was not realistic with the limited run time. The time limit also forced us to use 10x magnification for the Phikon model, but we expect it to improve with 20x magnification. We expect a CONCH model with 20x magnification to perform the best.

The competition data did, however, allow for a realistic situation regarding the data. The dataset included imperfect labels, as not all pathologists agree on a diagnosis. Furthermore, differences in imaging quality differed as well. Data from two different institutes was included, simulating real life scenarios where the amount of data increases model performance due to generalizability.

The current study failed to run a high amount of test runs on the test set, due to submission errors on Kaggle. One of the biggest problems that occurred is related to the disabled internet which was mandatory for submission. We found that the Trident model in its developed form needed internet to be usable. This led us to be forced to change the code internally, which did cost a lot of time.

Due to many setbacks, mostly related to submissions on Kaggle, we did not find the time to fully optimize our models. We propose future research to implement an optimized feature extraction method, which we expect to result in a better score.

V. CONCLUSION

The current research aims to show the advancements made in pathological analysis of prostate cancer in the past five years. For this we use data from the PANDAs 2020 competition, hosted on Kaggle. We aim to improve the highest leaderboard score from the competition in 2020, which was 0.94. To do this, we decided on a pipeline using Trident, Titan, and Phikon. Kaggle challenges do have limitations for submission, which included a 6 hour run time limit on GPU, a 9 hour run time limit on CPU, and internet had to be disabled for submission. Furthermore, the test set is hidden. This did limit the possibilities a fair amount.

We considered CONCH as foundation model, but this was not realistic with the limited run time. Besides this, 20x magnification is also expected to increase the results, however, this was unrealistic in regards to the run time limit for our pipeline. We decided on a Phikon foundational model with 10x magnification, as this was expected the yield the best results. However, our best performing model only yielded a score of 0.76. We expect future research to yield better results with optimized feature extraction methods. Furthermore, we expect CONCH to yield better results as well with similar settings. With the fast advancing field of machine learning, faster processing and less computationally expensive models are expects in the near future. Even though the models do not yet reach 100 percent accuracy, it is not to say whether this is not possible in a few years.

Acknowledgment

The authors would like to give a special thanks to prof. dr. Geert Litjens of the Radboud University Medical Center for his supervision and valuable insights during this research.

References

[1] Amelie Echle et al. "Deep learning in cancer pathology: a new generation of clinical biomarkers". In: *British Journal of Cancer* 124.4 (Feb. 16, 2021), pp. 686–696. ISSN: 0007-0920, 1532-1827. DOI: 10.1038/s41416-020-01122-x. URL: https://www.nature.com/articles/s41416-020-01122-x.

[2] Jana Lipkova et al. "Artificial intelligence for multimodal data integration in oncology". In: *Cancer cell* 40.10 (2022), pp. 1095–1110.

[3] Gregory Verghese et al. "Computational pathology in cancer diagnosis, prognosis, and prediction–present day and prospects". In: *The Journal of Pathology* 260.5 (2023), pp. 551–563.

[4] Haitham Kussaibi. "AI-Enhanced Subtyping of Thymic Tumors: Attention-based MIL with Pathology-Specific Feature Extraction". In: *medRxiv* (2024), pp. 2024–06.

[5] Alexandre Filiot et al. "Scaling self-supervised learning for histopathology with masked image modeling". In: *medRxiv* (2023), pp. 2023–07.

[6] Hamid Reza Tizhoosh and Liron Pantanowitz. "Artificial Intelligence and Digital Pathology: Challenges and Opportunities". In: *Journal of Pathology Informatics* 9.1 (Jan. 2018), p. 38. ISSN: 21533539. DOI: 10.4103/jpi.jpi_53_18. URL: https://linkinghub.elsevier.com/retrieve/pii/S2153353922003510.

[7] Ali Imran. *Scaling Self Supervised Learning for Histology*. URL: https://huggingface.co/blog/EazyAl/phikon.

[8] Andrew Zhang et al. "Accelerating Data Processing and Benchmarking of AI Models for Pathology". In: *arXiv preprint arXiv:2502.06750* (2025).

[9] Tong Ding et al. "Multimodal whole slide foundation model for pathology". In: *arXiv preprint arXiv:2411.19666* (2024).

[10] Geert Litjens et al. *Prostate cANcer graDe Assessment (PANDA) Challenge*. https://kaggle.com/competitions/prostate-cancer-grade-assessment. Kaggle. 2020.

[11] Nitin Singhal et al. "A deep learning system for prostate cancer diagnosis and grading in whole slide images of core needle biopsies". In: *Scientific Reports* 12.1 (Mar. 1, 2022), p. 3383. ISSN: 2045-2322. DOI: 10.1038/s41598-022-07217-0. URL: https://www.nature.com/articles/s41598-022-07217-0 (visited on 06/01/2025).

[12] Wouter Bulten et al. "Artificial intelligence for diagnosis and Gleason grading of prostate cancer: the PANDA challenge". In: *Nature Medicine* 28.1 (Jan. 2022), pp. 154–163. ISSN: 1078-8956, 1546-170X. DOI: 10.1038/s41591-021-01620-2. URL: https://www.nature.com/articles/s41591-021-01620-2.

[13] Jonathan I. Epstein et al. "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". In: *American Journal of Surgical Pathology* 40.2 (Feb. 2016), pp. 244–252. ISSN: 0147-5185. DOI: 10.1097/PAS.0000000000000530. URL: https://journals.lww.com/00000478-201602000-00010.

[14] PND. *PND Submission, Prostate cANcer graDe Assessment (PANDA) Challenge*. en. URL: https://kaggle.com/prostate-cancer-grade-assessment.

[15] Peter Neidlinger et al. "Benchmarking foundation models as feature extractors for weakly-supervised computational pathology". In: *arXiv preprint arXiv:2408.15823* (2024).