Evaluating Computational Pathology Foundation Models for Prostate Cancer Grading under Distribution Shifts

Fredrik K. Gustafsson
Karolinska Institutet
www.fregu856.com

The 30th Mayo-KI Annual Scientific Research Meeting
October 16, 2024

About Me

- Postdoctoral researcher in the group of Mattias Rantalainen at Karolinska Institutet, Department of Medical Epidemiology and Biostatistics.

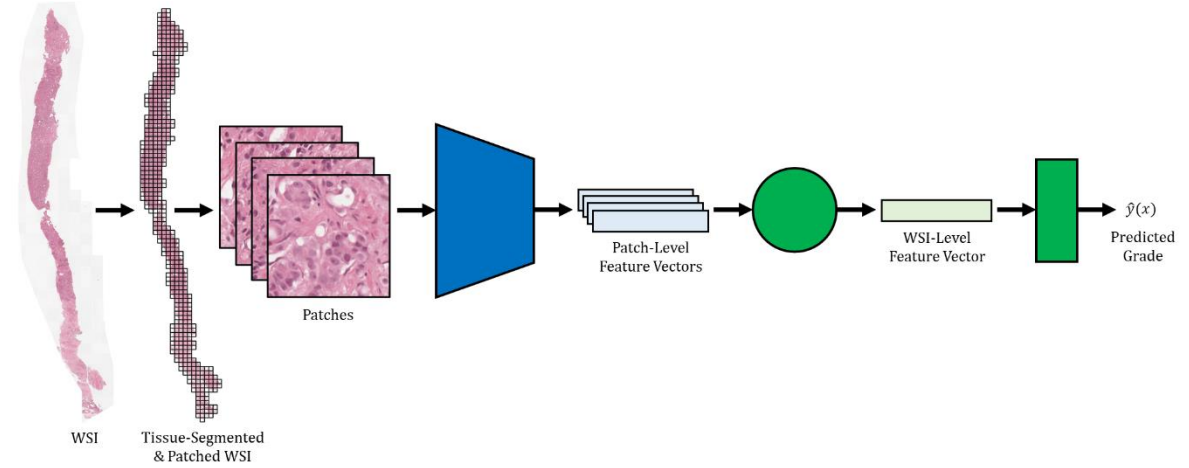
- Background:
  - → 2023: PhD in *Machine Learning*, Uppsala University.
  - → 2018: MSc in *Electrical Engineering*, Linköping University.
    - 2016–2017: Graduate exchange student, Stanford University.
  - → 2016: BSc in *Applied Physics and Electrical Engineering*, Linköping University.

- Machine learning and computer vision for *computational pathology*.

- My research focuses on how to build and evaluate *reliable machine learning* models, for applications within *data-driven medicine and healthcare*.

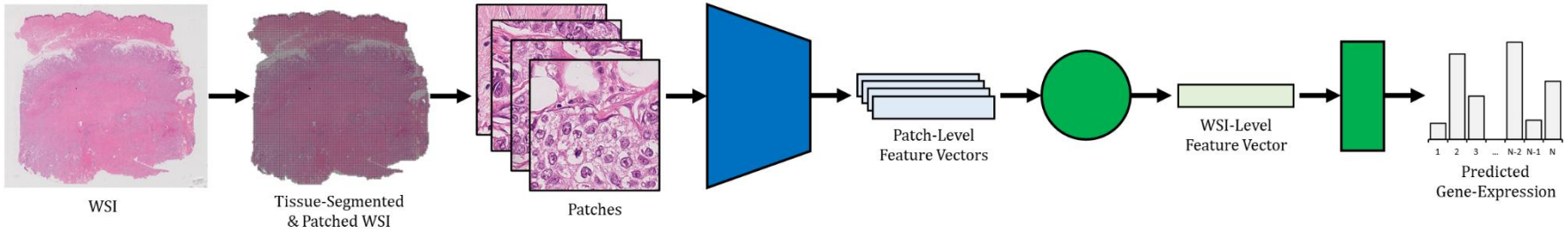
About the Presentation

Evaluating Computational Pathology Foundation Models for Prostate Cancer Grading under Distribution Shifts
Fredrik K. Gustafsson, Mattias Rantalainen, Preprint, 2024-10



Computational Pathology

- Computational pathology uses machine learning and computer vision to automatically extract useful information from histopathology whole-slide images (WSIs).
- Given datasets of (WSI, label) pairs, models can be trained for applications such as histological grading, risk stratification, prediction of biomarkers and gene-expression.



Foundation Models

- Foundation models are large deep learning models trained on large amounts of data using self-supervised learning.

- Have recently become a popular research direction within computational pathology.
  - → *Towards a General-Purpose Foundation Model for Computational Pathology* (Nature Medicine, 2024).
  - → *A Visual-Language Foundation Model for Computational Pathology* (Nature Medicine, 2024)
  - → *A Whole-Slide Foundation Model for Digital Pathology from Real-World Data* (Nature, 2024)
  - → *A Foundation Model for Clinical-Grade Computational Pathology and Rare Cancers Detection* (Nature Medicine, 2024)

- Foundation models are intended to be *general-purpose feature extractors*, promising to achieve good performance on a wide range of downstream prediction tasks.

Evaluated Foundation Models

- **UNI:** Vision-only foundation model.
  Towards a General-Purpose Foundation Model for Computational Pathology (Nature Medicine, 2024).
  - → Pretrained using self-supervised learning on a pan-cancer dataset (20 major tissue types) of roughly *100 million tissue patches* from more than *100,000 WSIs*.

- **CONCH:** Vision-language foundation model.
  A Visual-Language Foundation Model for Computational Pathology (Nature Medicine, 2024)
  - → First pretrained on a dataset of 16 million tissue patches from more than 21,000 WSIs.
  - → Then further pretrained using a vision-language objective on a dataset of more than *1.1 million image-caption pairs* (curated via processing of figures from PubMed articles).

- Resnet-IN: Baseline model.
  - → Pretrained on the ImageNet dataset of natural images.

Research Question

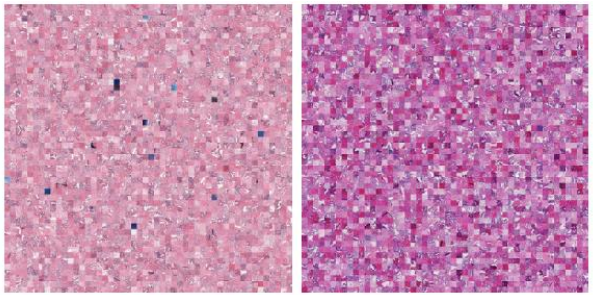
- Real-world pathology image data exhibits considerable variability, due to differences in staining and scanning procedures employed at different labs and hospitals.
- A truly general-purpose foundation model should be robust to these variations and other *distribution shifts* which might be encountered during practical deployment.
- While general deep learning models can be highly sensitive to distribution shifts, this has yet to be studied specifically for computational pathology foundation models.
- *Does the large and varied datasets utilized in the training of computational pathology foundation models make them robust to commonly encountered distribution shifts, or can the model performance still break down in certain practical settings?*

Prostate Cancer Grading

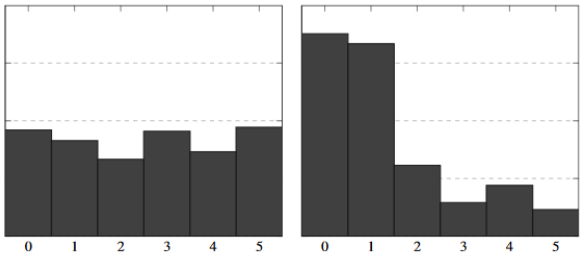
- Specific practical application: histological grading of prostate cancer biopsy WSIs.
- Histological grading is used to provide important prognostic information for patients, categorizing biopsies into five different ISUP grade groups.
  - → With non-tumor biopsies as grade 0, each biopsy WSI is assigned an ISUP grade 0 - 5.
- We conduct experiments on the publicly available PANDA dataset, containing more than 10,000 prostate biopsy WSIs with corresponding ISUP grade labels.
- The data was collected from two different sites: Radboud University Medical Center (Radboud) in the Netherlands, and Karolinska Institutet (Karolinska) in Sweden.

Evaluated Distribution Shifts

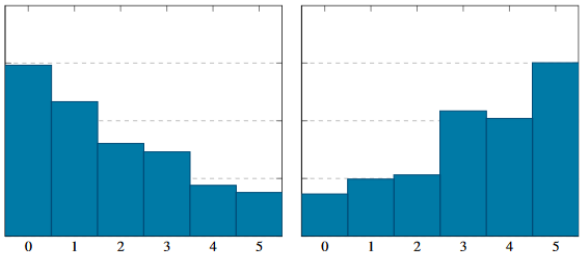
- Radboud and Karolinska differ in terms of both the pathology lab procedures and utilized scanners, creating a clear distribution shift for the WSI image data.
- By creating further subsets of the PANDA dataset, we are also able to evaluate robustness in terms of shifts in the label distribution over the ISUP grades 0 - 5.



(a) WSI image data shift, *Radboud* → *Karolinska*.



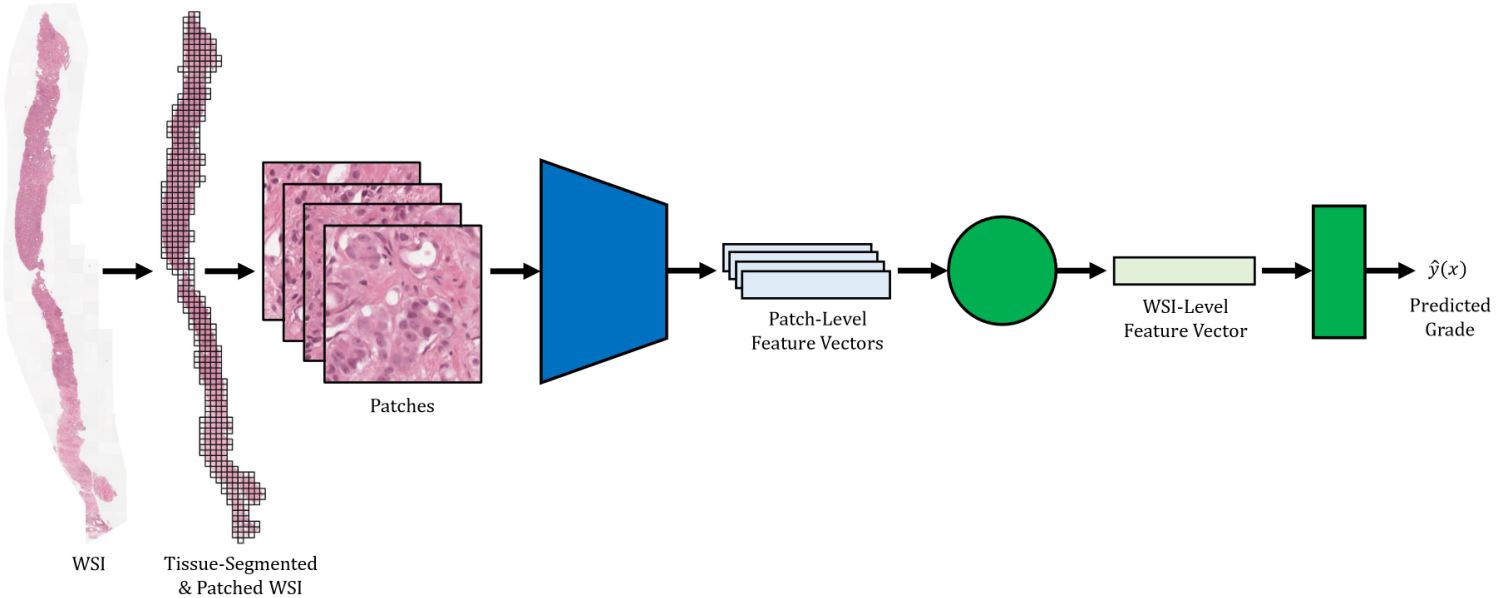
(b) Grade label shift, *Radboud* → *Karolinska*.



(c) Grade label shift, *Radboud-L* → *Radboud-R*.

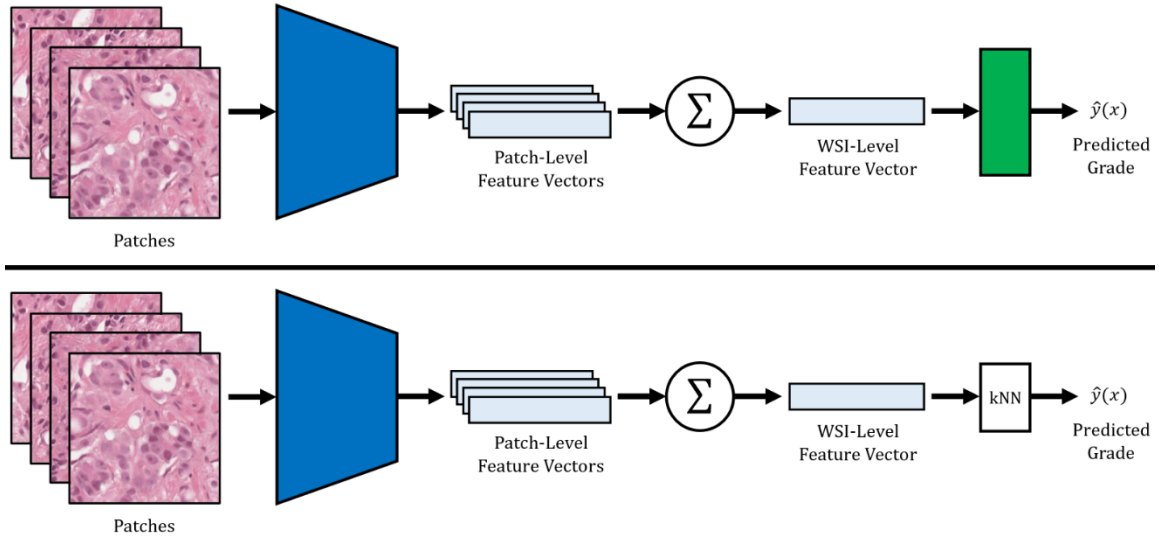
Application of Foundation Models – ABMIL

- We evaluate the foundation models by utilizing them as frozen *patch-level feature extractors* in three different ISUP grade prediction models.



Application of Foundation Models – Mean Feature, kNN

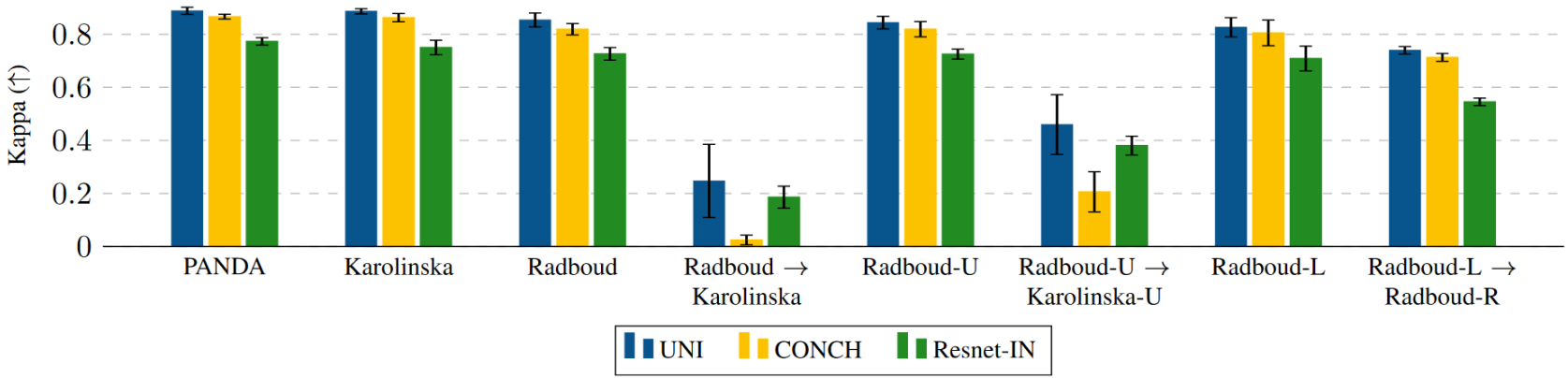
- We evaluate the foundation models by utilizing them as frozen *patch-level feature extractors* in three different ISUP grade prediction models.



Results (ABMIL)

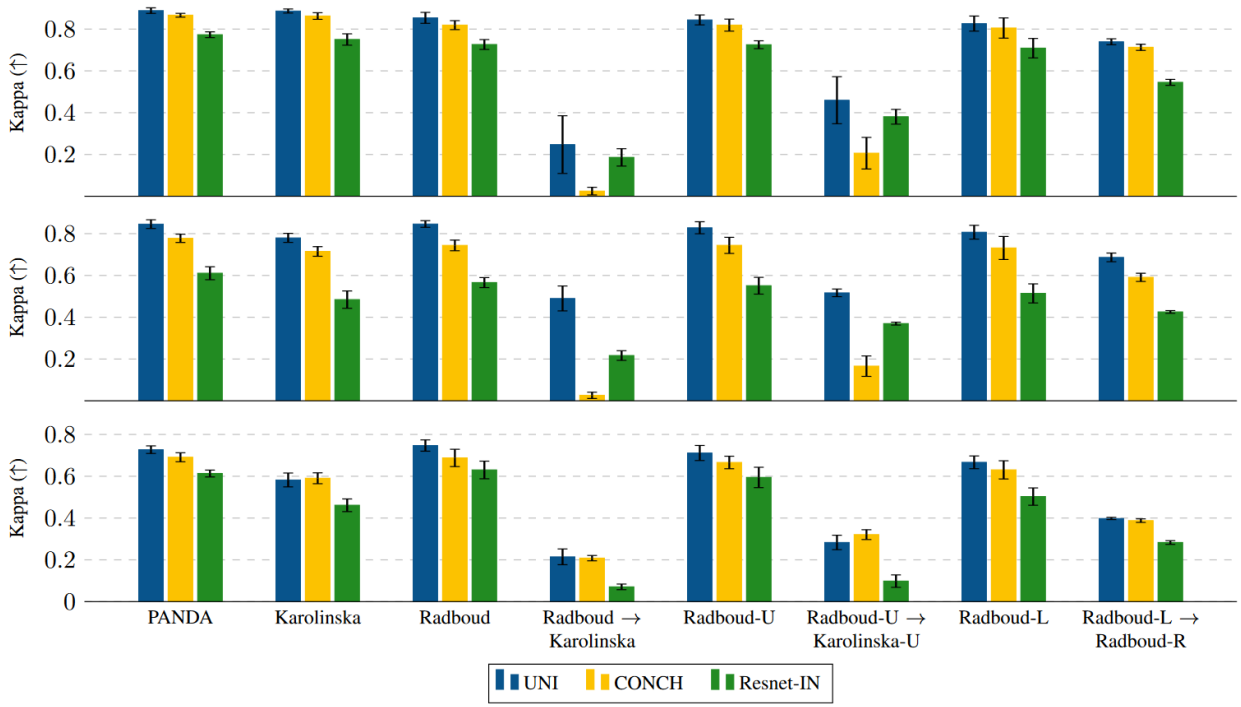
- When models are trained and evaluated on the full PANDA dataset, both UNI and CONCH perform well (0.89 kappa for UNI) and outperform Resnet-IN.
  - → Similar results are achieved also when models are both trained and evaluated exclusively on data from either Karolinska or Radboud.

- However, when models are trained on Radboud data and evaluated on Karolinska data, the performance drops drastically (0.25 kappa for UNI).



Results (ABMIL, Mean Feature, kNN)

- This clear performance drop is observed for all three ISUP grade models, i.e. even when applying kNN directly on top of the foundation model patch-level features.



Main Actionable Takeaways

- **(1/3)** While the computational pathology foundation models UNI and CONCH achieve very strong performance *relative* to the Resnet-IN baseline, the *absolute* performance can still be far from satisfactory in certain settings.

- **(2/3)** The fact that UNI and CONCH have been trained on very large and varied datasets does *not* guarantee that downstream prediction models always will be robust to commonly encountered distribution shifts.

- **(3/3)** Even within the paradigm of powerful pathology-specific foundation models, the quality of the data utilized to fit downstream prediction models is a crucial aspect.
  → If this data has limited variability (in terms of the number of data collection sites or utilized scanners), downstream models can still become sensitive to common distribution shifts.

Questions?

Fredrik K. Gustafsson
fredrik.gustafsson@ki.se
www.fregu856.com