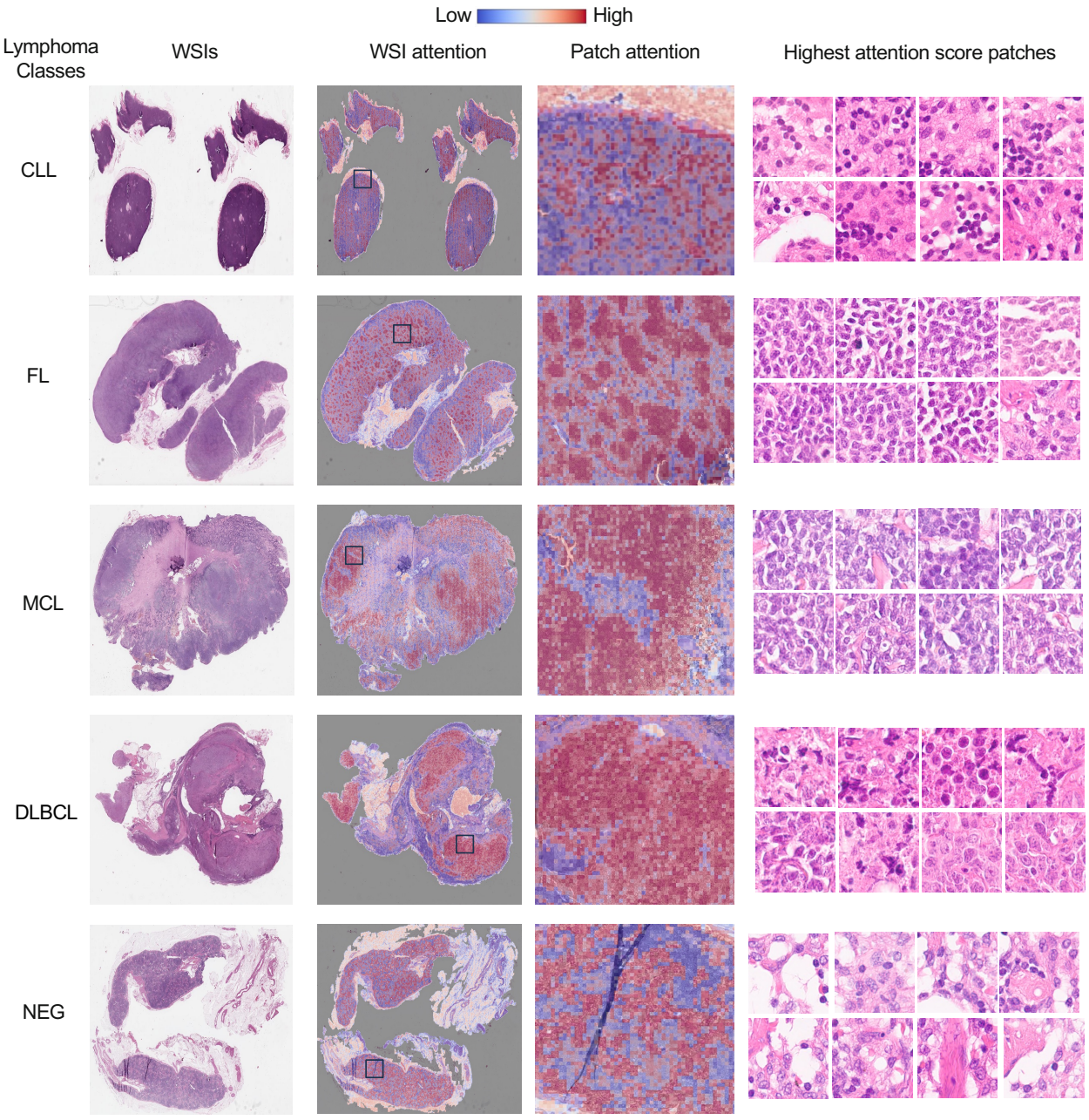
A Multicenter Benchmark of Multiple Instance Learning Models for Lymphoma Subtyping from HE-stained Whole Slide Images

Supplementary Material

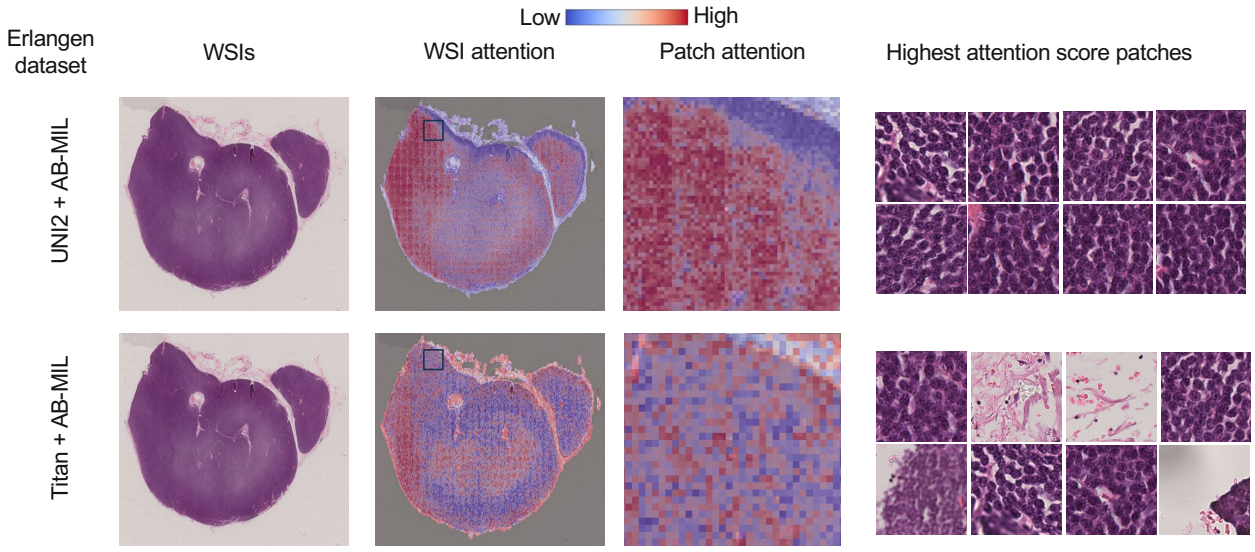
Rao Muhammad Umer[1] Daniel Sens[1] Jonathan Noll[1] Sohom Dey[1] Christian Matek[1,2,7] Lukas Wolfseher[8] Rainer Spang[8] Ralf Huss[9]
Johannes Raffler[9] Sarah Reinke[10] Ario Sadafi[1,4,5] Wolfram Klapper[10] Katja Steiger[6] Kristina Schwamborn[6] Carsten Marr[1,2,3,4]
[1] Institute of AI for Health, Helmholtz Munich, Neuherberg, Germany
[2] Department of Medicine III, Ludwig-Maximilian-University Hospital, Munich, Germany
[3] German Cancer Consortium (DKTK), Partner Site Munich, Germany
[4] Munich Center for Machine Learning (MCML), Munich, Germany
[5] Computer Aided Medical Procedures, Technical University of Munich, Munich, Germany
[6] Institute of Pathology, Technical University of Munich, School of Medicine and Health, Munich, Germany
[7] Institute of Pathology, Erlangen, Germany
[8] University of Regensburg, Regensburg, Germany
[9] Institute for Digital Medicine, University Hospital, Augsburg, Germany
[10] Department of Pathology, Hematopathology Section and Lymph Node Registry,
University Hospital Schleswig-Holstein, Kiel, Germany
carsten.marr@helmholtz-munich.de

1. Whole slide image attention visualization

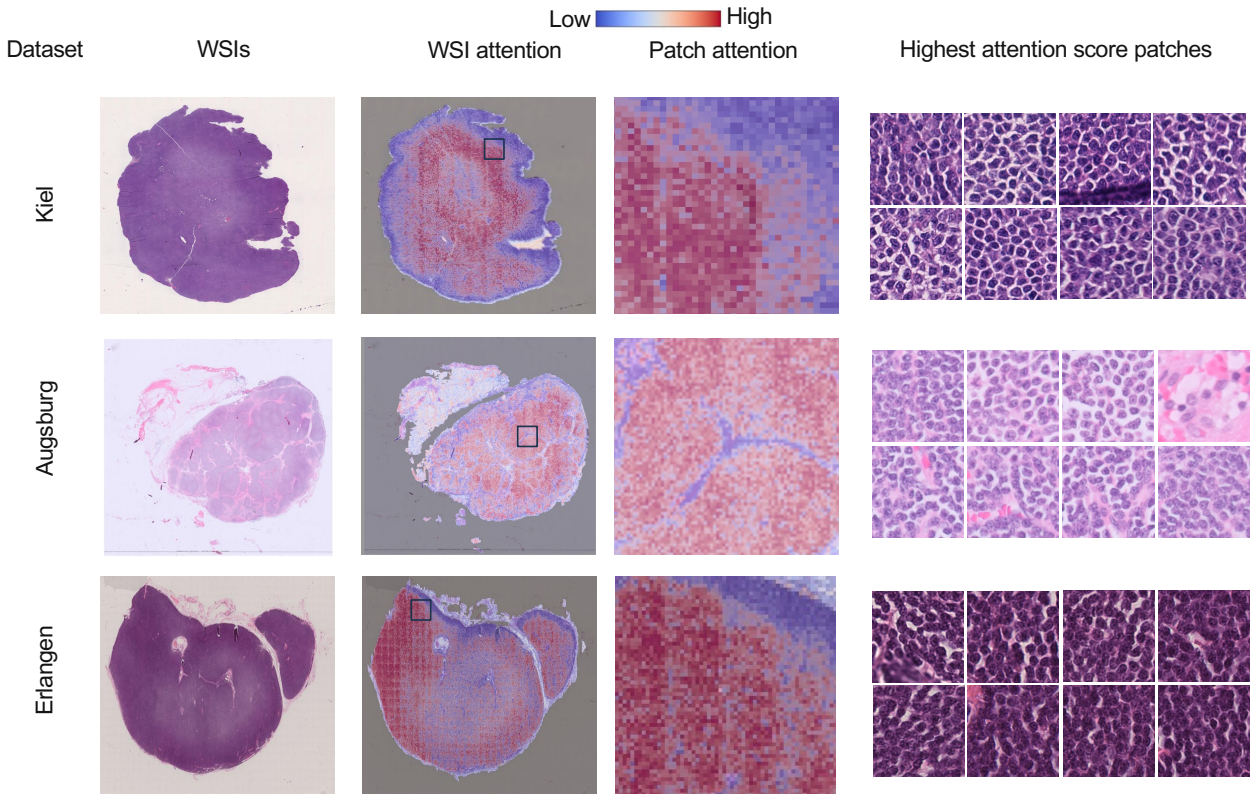
Attention scores can be visualized as heatmaps to highlight diagnostically informative regions, i.e., areas assigned high attention, while de-emphasizing regions of low relevance, such as normal tissue or background artifacts. To interpret the spatial distribution of model attention across a WSI, we convert the attention scores corresponding to the predicted class into percentiles and map these normalized values back to their original coordinates on the slide. Fine-grained heatmaps are produced by extracting overlapping patches and averaging the attention values within the overlapping areas. In our workflow, we employ Trident [24] to generate WSI-level attention visualizations (Supplementary Figures 1, 2, and 3) and to display the patches with the highest attention scores. In Supplementary Figure 2, UNI2 features generalize more robustly than Titan, and avoid artifact patches. We note that a high-attention score in CLL is given to non-lymphoma areas and in DLBCL to poorly preserved areas (Supplementary Figure 1). This behavior might be avoidable when training the algorithm with more cases.



Supplementary Figure 1. Attention visualization on IID Munich testset.



Supplementary Figure 2. Attention visualization of a CLL lymphoma WSI from the OOD Erlangen cohort. UNI2 avoids artifact patches with the highest attention scores as compared to Titan.



Supplementary Figure 3. Attention visualization of CLL lymphoma subtype on OOD lymphoma testsets.