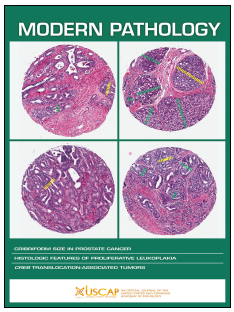
Journal Pre-proof



Fast-track development and multi-institutional clinical validation of an artificial intelligence algorithm for detection of lymph node metastasis in colorectal cancer

Avri Giammanco, Andrey Bychkov, Simon Schallenberg, Tsvetan Tsvetkov, Junya Fukuoka, Alexey Pryalukhin, Fabian Mairinger, Alexander Seper, Wolfgang Hulla, Sebastian Klein, Alexander Quaas, Reinhard Büttner, Yuri Tolkach

PII: S0893-3952(24)00076-0

DOI: https://doi.org/10.1016/j.modpat.2024.100496

Reference: MODPAT 100496

To appear in: *Modern Pathology*

Received Date: 23 December 2023

Revised Date: 24 March 2024

Accepted Date: 10 April 2024

Please cite this article as: Giammanco A, Bychkov A, Schallenberg S, Tsvetkov T, Fukuoka J, Pryalukhin A, Mairinger F, Seper A, Hulla W, Klein S, Quaas A, Büttner R, Tolkach Y, Fast-track development and multi-institutional clinical validation of an artificial intelligence algorithm for detection of lymph node metastasis in colorectal cancer, *Modern Pathology* (2024), doi: https://doi.org/10.1016/j.modpat.2024.100496.

Fast-track development and multi-institutional clinical validation of an artificial intelligence algorithm for detection of lymph node metastasis in colorectal cancer

Avri Giammanco[1], Andrey Bychkov[2,3], Simon Schallenberg[4], Tsvetan Tsvetkov[1], Junya Fukuoka[2,3], Alexey Pryalukhin[5], Fabian Mairinger[6], Alexander Seper[5,7], Wolfgang Hulla[5], Sebastian Klein[1], Alexander Quaas[1], Reinhard Büttner[1], Yuri Tolkach[1#]

[1]Institute of Pathology, University Hospital Cologne, Cologne, Germany

[2]Department of Pathology, Kameda Medical Center, Kamogawa, Japan

[3]Department of Pathology Informatics, Nagasaki University, Nagasaki, Japan

[4]Institute of Pathology, Charite University Clinic, Berlin, Germany

[5]Institute of Pathology, Wiener Neustadt State Hospital, Wiener Neustadt, Austria

[6]Institute of Pathology, University Hospital Essen, Essen, Germany

[7]Danube Private University, Wien, Austria

#Correspondence:

Yuri Tolkach MD PhD

Institute of Pathology

University Hospital Cologne

Cologne, Germany

Phone: +49 221 478 6365

Email: yuri.tolkach@gmail.com

Abstract

Lymph node metastasis (LNM) detection can be automated using artificial intelligence-based diagnostic tools. Only limited studies have addressed this task for colorectal cancer. The aim of this study was to develop of a clinical-grade digital pathology tool for LNM detection in colorectal cancer (CRC) using the original fast-track framework.

The training cohort included 432 slides from one department. A segmentation algorithm detecting 8 relevant tissue classes was trained. The test cohorts consisted of materials from five pathology departments digitized by four different scanning systems.

A high-quality, large training dataset was generated within 7 days, and a minimal amount of annotation work using fast-track principles. The AI tool showed very high accuracy for LNM detection in all cohorts, with sensitivity, negative predictive value, and specificity ranges of 0.980-1.000, 0.997-1.000, and 0.913-0.990, correspondingly. Only 5 of 14460 analyzed test slides with tumor cells over all cohorts were classified as false negative (3/5 representing clusters of tumor cells in lymphatic vessels).

A clinical-grade tool was trained in a short time using fast-track development principles and validated using the largest international, multi-institutional, multi-scanner cohort of cases to date, showing very high precision for LNM detection in CRC. We are releasing a part of the test datasets to facilitate academic research.

Key words:

Lymph node, metastasis detection, AI, colorectal cancer, validation, digital pathology.

Introduction

Recent developments in digital and computational pathology are transforming our approaches to diagnostics within the field of pathology[1,2]. Numerous studies have highlighted the promising capabilities of artificial intellegence (AI) algorithms applied to common pathology diagnostic tasks, such as tumor detection, histological grading, subtyping, regression grading, and many others[3–8].

Lymph node metastasis detection is one of the most common and time-consuming manual tasks pathologists perform during the processing of lymphadenectomy specimens. Numerous slides must be analyzed, and all metastases, regardless of size, must be detected, which is crucial for cancer reporting, staging, and risk stratification. This area naturally lends itself to the application of diagnostic AI algorithms and has garnered early attention from computational pathology researchers. One of the most notable efforts to stimulate the development of lymph node metastasis detection tools was the CAMELYON16/17 Challenge[9–11]. This challenge addressed the most critical issue in AI algorithm development, which remains relevant today: the absence of high-quality training datasets.

Although breast cancer metastasis detection has received considerable attention due to the CAMELYON challenge, lymph node metastasis detection in other tumors remains understudied. Thus, only a few studies have addressed this task for colorectal cancer[12–16], with only one study[15] using a small external cohort to validate the results, which is of utmost importance for diagnostic AI tools. All of these studies utilized classification neural networks, analyzing slides in regions (tiles), rather than pixel-wise, which is outdated for a diagnostic tool in pathology.

In this study, we developed a clinical-grade, precise segmentation computational pathology tool for the detection of lymph node metastasis in colorectal cancer. To

achieve this, we introduced and utilized the fast-track development framework, which addresses the major challenge of computational pathology – the creation of training datasets. We demonstrate how a high-quality, large training dataset can be created in less than one week with minimal expert time spent on annotations, utilizing pre-existing tools. Finally, we validated our algorithm using the largest international external cohort of cases to date, consisting of thousands of slides from five pathology departments scanned by four different scanning systems approved for diagnostic use. Our algorithm exhibited very high sensitivity (1.00 sensitivity in 5 out of 7 cohorts) and specificity at a high analysis speed of single slides. Furthermore, we have publicly released part of our test datasets to facilitate research in this domain.

Materials and Methods

Training cohort

Cologne, Germany (UKK), and were retrospectively identified from archived pathology cases of patients who underwent surgery between 2017 and 2019. These cases represented all stages, morphologic variants, and histological grades and contained only colorectal adenocarcinoma. Both benign and tumor-containing lymph node slides were identified by a board-certified pathologist for inclusion in the training dataset. Additionally, 61 archive slides representing only non-specific inflammatory altered/necrotic fatty tissue and associated resorption were included to enrich the training dataset for inflamed fatty tissue often seen in lymph nodes residing near the primary tumor. Furthermore, a round of hard-negative mining was carried out on 111 additional slides from the same cases (but non-annotated) included in the training dataset, automatically extracting and adding difficult regions

from the slides into the training dataset. Next, to further increase the accuracy of the classifier, lymph node classifier-relevant classes were extracted from our primary tumor dataset, published previously[5].

Test cohorts

Archive lymphadenectomy specimens accompanying primary colorectal tumor resections were collected from five pathology departments: University Hospital Cologne (UKK; Germany; 2021-2022, using the concept of temporal validation, at least 3 years older than slides from the training dataset), State Hospital Wiener Neustadt (WNS; Austria), Charité University Hospital (CHA; Germany), Kameda Medical Center (KAM; Japan), and University Hospital Essen (ESS; Germany). All series represented consecutive cases from routine diagnostics without any preselection, encompassing all grades, stages, and morphologies of colorectal adenocarcinoma. Manual and automated quality control was performed after digitization, and slides with large out-of-focus artifacts were excluded from the final analysis. Some slides could not be digitized by scanners due to mechanical defects or other reasons. These two factors accounted for differences in the number of slides in cohorts (UKK, WNS) digitized by several scanners.

Digitization

The training dataset from UKK was digitized using a Hamamatsu NanoZoomer S360 scanner (Hamamatsu Photonics, Japan). The training dataset of primary tumors from a previous publication was digitized using multiple Leica Aperio scanners[5]. The test datasets from UKK and WNS were digitized with two scanners each (Hamamatsu NanoZoomer S360 and Leica Aperio GT450, Leica, Wetzlar,

Germany) to account for heterogeneity due to the use of different scanning systems. The KAM test dataset was digitized using a Philips UltraFast scanner (Philips, Netherlands). The ESS test dataset was digitized using a 3DHISTECH Pannoramic 250 scanner. For further details including slide micron-per-pixel (MPP) parameter, see Figure 1B.

Annotation principles

All annotations were performed in QuPath version 0.3.2. For pre-annotation purposes, representative rectangular regions from tumor-containing slides were selected by an experienced pathologist (YT) in QuPath and automatically extracted as JPEG files for all slides. These regions were then processed to create binary maps for all classes, except benign lymph node tissue, using an algorithm published elsewhere, outside of QuPath. These binary maps were automatically imported into QuPath, resulting in pre-annotations. These pre-annotations were then corrected by experienced human analysts (AG, YT). Annotations for benign lymph node tissue were created using the Thresholder function in QuPath, with parameters adapted to achieve better pre-annotation quality for individual slides.

Algorithm development and training

The algorithm development was conducted using the PyTorch framework version 1.13 and Python version 3.9. For constructing the algorithm's segmentation models, the PyTorch package version 0.3.1 was employed. The final version of the algorithm is based on the UNet++ neural network architecture and utilizes the EfficientNetB0 encoder pre-trained on ImageNet, serving as a segmentation network for pixel-wise classification of whole-slide images. Training was carried out using NVIDIA A100 and

V100 graphics processing units (NVIDIA, Santa Clara, CA, USA). A small validation subset of the training dataset (approximately 10%) was reserved for validation purposes and algorithm fine-tuning. Training involved balanced sampling of all classes within batches and oversampling of under-represented classes for up to 5 times during single epochs. Extensive data augmentation techniques were employed during training, including flips, rotations, adjustments to brightness, contrast, gamma, hue, and saturation, following the approach outlined by Tellez et al[17]. No stain normalization was used at any development steps. The final version of the algorithm analyses images at the resolution 1.0 µm per pixel (patch size 512 pixels).

Clinical validation

For clinical validation purposes, each distinct lymph node section in the whole slide image was considered as a separate lymph node section to enable slide-level metrics. All the slides included were reviewed centrally once again by a board-certified pathologist (YT). Technically, the images were read at the original resolution (whereby original resolution/MPP is known and can be extracted from the metadata) and the patches for processing with the algorithm (512x512 px at MPP 1.0) were extracted under original resolution of the slide (patch size under original resolution = MPP model / MPP original * 512) and then simply resized to 512 px.

Statistical analysis

All statistical tests for lymph node metastasis detection were conducted at the slide level using R version 4.1.3 (The R Foundation for Statistical Computing). Typical metrics were calculated, including specificity, sensitivity, negative and positive predictive values, overall accuracy, and F1 score. A post-hoc analysis of different

thresholds for slide classification was performed after the initial validation was completed.

Results

Fast-track algorithm development

In this study, we applied several principles that enabled fast-track algorithm development, specifically for the creation of high-quality annotated training data, which was accomplished in approximately one week (the full pipeline is presented in Figure 1C). For comparison, the preparation of a high-quality training dataset typically takes 10-12 months[5]. Initially, we generated pre-annotations using a pre-trained algorithm from the same domain (but for primary tumors, see Methods) capable of recognizing 7 out of 8 classes necessary for the lymph node metastasis detection task but in the context of primary tumors[5]. These classes show almost no visual differences in the context of metastases, which allowed for the generation of precise pre-annotations with only a need for minor corrections to achieve very high-quality pixel-wise annotations. The only tissue class not available in the pre-trained algorithm is benign lymph node tissue, which was then automatically annotated using QuPath's native capabilities, resulting in very precise annotations that were minimally corrected by human analysts to remove any inconsistencies. To further improve accuracy, we enriched our dataset with inflamed adipose tissue (Figure 1C) and finally included regions with relevant classes from our primary tumor dataset (Figure 1C).

These steps resulted in the training of a precise segmentation algorithm in a very short time. The principle of the algorithm is intuitive (Figure 2); it analyzes the tissue sections, detects different tissue classes, and identifies tumor tissue corresponding to metastasis. Pathologists are presented with a color map overlaid on the original

whole-slide image, allowing for easy identification of regions in the slide with a high possibility of being a tumor (Figure 2).

Clinical validation of the algorithm

To validate the algorithm, we created a large, multi-institutional cohort of lymph node slides consisting of consecutive mesenteric lymphadenectomy cases without pre-selection from five departments across three countries. This cohort encompassed 8,967 lymph node sections (Figure 1B). The slides were digitized using four different scanning systems (Figure 1B, Figure 3A), with two of them (UKK, WNS) digitized by two distinct scanning systems. The validation results at the slide level demonstrated very high sensitivity (0.980-1.000) and negative predictive value (NPV, 0.997-1.000) for the algorithm, along with similarly high specificity (0.913-0.990) across all tested cohorts (Figure 3B). In five out of seven tested cohorts, all tumor slides were accurately detected, resulting in a sensitivity and NPV of 1.000. We conducted an additional post-hoc evaluation of the different thresholds for areas with a high probability of being tumor tissue and its effect on the final slide classification across all cohorts (see Suppl. Fig. 1 for cohort-level metrics and Suppl. Table 1 for slide-level metrics). This analysis demonstrates that the initially selected, very small cutoff is optimal for the high sensitivity of the algorithm.

Detailed evaluation of false negative and false positive detections

Only individual slides with tumors in the CHA and KAM cohorts were not classified as such, due to an initially selected, very small area cutoff for whole-slide image classification (Figure 4C; corresponding to approximately the area of 5 tumor cells, see Methods). There were no false-negative detections in the UKK, WNS, and ESS

cohorts. Among these individual slides in CHA and KAM, the tumor was detected in all these slides (Figure 5A-D) and would have been presented to the pathologist for visual evaluation. Three out of five of these regions were tumor cell clusters in sinuses without invasion, which we still considered as lymph node metastases in our metrics.

All false positive results were reasonable. There were no misclassifications where the underlying algorithm decision was not understandable. Most false positive regions were small and served as valuable alerts that could enhance the confidence of the diagnostic process (Figure 6A-C). Such areas included hyperactivated histiocytes, where immunohistochemical clarification might be necessary, and severe mechanical artifacts (crush artifacts) where the tissue was not evaluable and might conceal metastasis, as well as regions of fibrosis suspicious of desmoplastic tumor stroma. Occasionally, blood vessels were detected as areas with a high probability of being a tumor. In the KAM cohort, the algorithm produced slightly more false positive misclassifications related to activated germinal centers (Figure 6F).

Time-to-analysis metrics

The algorithm facilitated rapid analysis of whole slide images. The average (median) processing time per lymph node section ranged from 23 to 50 (18 to 30) seconds, enabling efficient processing of the entire case through parallelization. For instance, a consumer-grade GPU unit, such as the RTX 4090, can process up to 10 whole slide images simultaneously. Further details are provided in Figure 4D.

Discussion

Developing clinical-grade assistive AI tools for diagnostic pathology presents significant challenges for several reasons. First, creating a robust diagnostic tool that can generalize to unseen data requires large, diverse training datasets with highly accurate annotations. Securing the necessary volume of high-quality annotations is a laborious process that involves expert pathologists and can take several months of meticulous effort. Furthermore, the tool must undergo rigorous clinical validation with previously unseen cases to prove its reliability and effectiveness.

In this study, we introduce our fast-track development principle that capitalizes on existing domain knowledge and previous advancements (Figure 1C). Utilizing this principle, we were able to develop a highly effective, clinical-grade diagnostic algorithm for detecting lymph node metastasis in colorectal cancer in a remarkably short period (Figure 1C, Figure 2). Our fast-track method required less than 7 days of hands-on time to compile a high-quality training dataset with highly precise annotations. For context, the development of a recently reported accurate, clinical-grade tool for processing primary colorectal cancer specimens took 16 months of annotation work[5]. This approach holds potential for application across various other domains.

We validated our algorithm using a multi-institutional, international cohort (Figure 1B, Figure 3A) comprised of consecutive lymphadenectomy specimens from five pathology departments (three in Germany, one in Austria, and one in Japan), digitized with the four most common scanning systems[18]. This continuous cohort, without pre-selection, captures cases of varying complexity, quality, and morphology. Our cohort is among the largest to date in published research. Our algorithm demonstrated very high sensitivity (0.980-1.000; with five out of seven cohorts showing a sensitivity of 1.000), negative predictive value (0.997-1.000), and

specificity (0.913-0.990), regardless of the cohort, digitization system, or cancer morphology (including mucinous), as illustrated in Figures 2 and 3. A detailed analysis of the few false negative results (Figure 4,5; five false negative misclassifications out of 14,460 analyzed lymph node sections) revealed that almost all tumor deposits were correctly detected. The misclassifications were due to a pre-defined, very small area threshold for positive slide classification. Three of the five "false negative" results were clusters of tumor cells in lymph vessels without invasion (Figure 5).

Interestingly, in two test cohorts scanned by the Hamamatsu scanner, a slightly lower specificity was observed (Figure 4B, cohorts WNS A and UKK B), underscoring the importance of incorporating various scanning systems into validation studies. Based on our experience, Hamamatsu scanners generate images with a darker color scheme compared to other scanners. In our specific task of metastasis detection, for the algorithm, "dark" could potentially serve as an additional signal leaning towards the "tumor" class. In our study, these slight drops in specificity were deemed acceptable.

Although it impacted the specificity metric, the majority of false positive misclassifications represented very useful alerts that drew the attention of pathologists, ensuring that all suspicious areas were thoroughly investigated. Most of these misclassifications involved severely artifactually altered lymph node tissue and over-activated histiocytes (Figure 6). This issue can be mitigated by enriching the training dataset with such cases.

Few recent studies have addressed the detection of colorectal cancer metastasis in lymph nodes[12,14–16,19]. Tan et al.[16] tackled the problem of lacking annotations through a Multiple Instance Learning (MIL) approach. MIL-based methods produce

algorithms that allow only regional (so-called "patch-level") classification, not pixel-wise, precise segmentation maps, which are outdated for diagnostic tools. The authors demonstrated an accuracy of 0.953 at slide-level classification for the training cohort but did not perform external validation. Chuang et al.[14] used a weakly-supervised approach similar to MIL with only slide-level labels (tumor/benign). However, both development and testing were performed on a dataset from a single department. In this dataset, the authors detected tumors through the generation of so-called class activation maps (pseudo-segmentation) and received an area under the receiver operating curve (AUROC) parameter of 0.9476-0.9944, depending on metastasis size. Kindler et al.12 developed a segmentation algorithm based on pixel-level annotations, which was then tested using slides from the same department (288 whole slide images covering 1,517 lymph node sections, of those with tumor 103). The authors reported a sensitivity of 0.990 on the slide level for the internal cohort.

A recent study by Khan et al. presents the most advanced study to date regarding external validation. The authors used one external cohort (1,033 slides) and three internal cohorts (2,803, 172, and 217 slides), achieving a sensitivity of 0.872-1.000 for tumor detection. The lower bound of this range is substantially lower than that of our model (0.872 vs. our 0.980). The algorithm of Khan et al. [15] is an ensemble of the Xception network and Vision Transformer, which may imply high computational costs and longer analysis times. This algorithm is also a patch-level classification that produces relatively rough pseudo-segmentation masks of tumor and tumor-associated tissue (tumor, stroma, mucin) as a single class. Our algorithm separately detects all tumor-related tissue classes, which is crucial, as necrosis, mucin, and tumor stroma/suspicious fibrosis may appear outside of the tumor tissue concept and

in the perilymphatic tissue, potentially resulting in false positive misclassifications. Our algorithm is resistant to this source of false positives. Moreover, it is a shallow network with very quick analysis times, averaging approximately 30 seconds per lymph node section (Figure 3D). Using external validation is of utmost importance for diagnostic tools, as poor generalization and overfitting to the training dataset is a common issue. In terms of external validation, our study is superior by a margin to the aforementioned studies (excluding UKK B and UKK C, four large external cohorts scanned by four different scanning systems).

Another study by Bandi et al.[19] deserves mention for its similarities to our fast-track principle, specifically its use of the continuous learning concept to decrease the number of annotations necessary for training the algorithm. This is achieved by transferring domain knowledge and materials from lymph node metastasis detection tasks for other tumors, such as breast and head-neck cancer. However, the aim of this study is to serve as a proof-of-concept for continual learning. The authors utilized a small dataset of 119 slides for colorectal cancer from a single department and demonstrated metrics that were inferior to those of our study.

Our study is not without limitations. One such limitation is the retrospective analysis of the cohorts. The algorithm should be validated prospectively and integrated into the diagnostic routines of pathology departments. This requires digitizing the sign-out process upfront. The principle of the algorithm dictates running it immediately after scanning the slides, ensuring that pathologists have all necessary outputs by the time diagnostics begin. This approach should save time, particularly in departments where colorectal resection specimens constitute a major part of the diagnostic workload, given the algorithm's very high sensitivity levels.

In conclusion, in this study, we propose fast-track principles for algorithm development by leveraging domain knowledge and previously collected data. Utilizing this approach, we demonstrate the creation of a high-quality training dataset in less than 7 days and the training of a clinical-grade, precise algorithm for detecting lymph node metastasis in colorectal cancer within a short timeframe. Our algorithm was validated using the largest multi-institutional, international dataset to date, comprising thousands of lymph node sections. It exhibits very high, clinical-grade sensitivity and specificity for metastasis detection. Additionally, we are releasing part of our test datasets to facilitate further research.

AI-assisted technologies usage statement

During the preparation of this work the authors used GPT4 and Grammarly in order to improve readability and language. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Author Contribution Statement

AG: data annotations, data management, data analysis, manuscript drafting; SS, AB, SS, JF, FM, AS, AP, WH: provided study data/cohorts, data preparation and management; TT: correction of pre-annotations, data management; SK, AQ: data analysis; RB: supervision, resources, data management and analysis; YT: conceptualization and design, annotations, technical development, algorithm training, data analysis, statistical analysis, supervision, resources, manuscript drafting. All authors: critical revision for important intellectual context.

Funding information

This project was funded by North Rhine-Westphalia state (European Fund for Regional Development (EFRE), 2014-2020; REACT-EU): Project DIGI-PATH (YT, RB), by Federal Ministry of Education and Research of Germany: Project FED-PATH (YT, RB, FM), and the Wilhelm Sander Foundation, Munich, Germany: Grant 2022.040.1 (YT). YT and AG have access to all the data.

Data availability

Upon publication, we will release a subset of our test datasets, both with and without metastasis, exclusively for academic research purposes on Zenodo. For access to the extended version of these datasets, the corresponding author should be contacted.

Ethics approval

Ethical committee approval for the use of retrospective archive material was obtained from all departments (UKK/CHA/ESS: joint 20-1583, WNS: GS1-EK-4/694-2021, KAM: 22-094). The requirement for specific patient permission was waived owing to the anonymized, retrospective nature of the material used.

Acknowledgments

We thank the Regional Computing Center of the University of Cologne (RRZK) for providing computing time on the DFG-funded (Funding number: INST 216/512/1FUGG) High Performance Computing (HPC) system CHEOPS as well as technical support.

Competing interests

All authors report no relevant conflicts of interest related to this study.

References

1. Bera K, Schalper KA, Rimm DL, Velcheti V, Madabhushi A. Artificial intelligence in digital pathology — new tools for diagnosis and precision oncology. *Nat Rev Clin Oncol*. 2019;16(11):703-715. doi:10.1038/s41571-019-0252-y
2. Echle A, Rindtorff NT, Brinker TJ, Luedde T, Pearson AT, Kather JN. Deep learning in cancer pathology: a new generation of clinical biomarkers. *Br J Cancer*. 2021;124(4):686-696. doi:10.1038/S41416-020-01122-X
3. Campanella G, Hanna MG, Geneslaw L, et al. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nat Med*. 2019;25(8):1301-1309. doi:10.1038/s41591-019-0508-1
4. Tolkach Y, Dohmgörgen T, Toma M, Kristiansen G. High-accuracy prostate cancer pathology using deep learning. *Nat Mach Intell*. 2020;2(7):411-418. doi:10.1038/s42256-020-0200-7
5. Griem J, Eich ML, Schallenberg S, et al. Artificial Intelligence-Based Tool for Tumor Detection and Quantitative Tissue Analysis in Colorectal Specimens. *Mod Pathol*. 2023;36(12):100327. doi:10.1016/J.MODPAT.2023.100327
6. Tolkach Y, Wolgast LM, Damanakis A, et al. Artificial intelligence for tumour tissue detection and histological regression grading in oesophageal adenocarcinomas: a retrospective algorithm development and validation study. *Lancet Digit Health*. 2023;5(5):e265-e275. doi:10.1016/S2589-7500(23)00027-4
7. Pantanowitz L, Quiroga-Garza GM, Bien L, et al. An artificial intelligence algorithm for prostate cancer diagnosis in whole slide images of core needle biopsies: a blinded clinical validation and deployment study. *Lancet Digit Health*. 2020;2(8):e407-e416. doi:10.1016/S2589-7500(20)30159-X
8. Bulten W, Pinckaers H, van Boven H, et al. Automated deep-learning system for Gleason grading of prostate cancer using biopsies: a diagnostic study. *Lancet Oncol*. 2020;21(2):233-241. doi:10.1016/S1470-2045(19)30739-9
9. Bándi P, Geessink O, Manson Q, et al. From Detection of Individual Metastases to Classification of Lymph Node Status at the Patient Level: The CAMELYON17 Challenge. *IEEE Trans Med Imaging*. 2019;38(2):550-560. doi:10.1109/TMI.2018.2867350
10. Litjens G, Bandi P, Bejnordi BE, et al. 1399 H&E-stained sentinel lymph node sections of breast cancer patients: the CAMELYON dataset. *Gigascience*. 2018;7(6):1-8. doi:10.1093/GIGASCIENCE/GIY065
11. Bejnordi BE, Veta M, Van Diest PJ, et al. Diagnostic Assessment of Deep Learning Algorithms for Detection of Lymph Node Metastases in Women With Breast Cancer. *JAMA*. 2017;318(22):2199-2210. doi:10.1001/JAMA.2017.14585
12. Kindler C, Elfwing S, Öhrvik J, Nikberg M. A Deep Neural Network-Based Decision Support Tool for the Detection of Lymph Node Metastases in Colorectal Cancer Specimens. *Mod Pathol*. 2023;36(2):100015. doi:10.1016/J.MODPAT.2022.100015
13. Davri A, Birbas E, Kanavos T, et al. Deep Learning on Histopathological Images for Colorectal Cancer Diagnosis: A Systematic Review. *Diagnostics*

2022, Vol 12, Page 837. 2022;12(4):837. doi:10.3390/DIAGNOSTICS12040837

14. Chuang WY, Chen CC, Yu WH, et al. Identification of nodal micrometastasis in colorectal cancer using deep learning on annotation-free whole-slide images. *Mod Pathol*. 2021;34(10):1901-1911. doi:10.1038/S41379-021-00838-2
15. Khan A, Brouwer N, Blank A, et al. Computer-Assisted Diagnosis of Lymph Node Metastases in Colorectal Cancers Using Transfer Learning With an Ensemble Model. *Modern Pathology*. 2023;36(5):100118. doi:10.1016/J.MODPAT.2023.100118
16. Tan L, Li H, Yu J, et al. Colorectal cancer lymph node metastasis prediction with weakly supervised transformer-based multi-instance learning. *Med Biol Eng Comput*. 2023;61(6):1565-1580. doi:10.1007/S11517-023-02799-X
17. Tellez D, Litjens G, Bándi P, et al. Quantifying the effects of data augmentation and stain color normalization in convolutional neural networks for computational pathology. *Med Image Anal*. 2019;58. doi:10.1016/j.media.2019.101544
18. Pinto DG, Bychkov A, Tsuyama N, Fukuoka J, Eloy C. Real-World Implementation of Digital Pathology: Results From an Intercontinental Survey. *Laboratory Investigation*. 2023;103(12):100261. doi:10.1016/J.LABINV.2023.100261
19. Bándi P, Balkenhol M, van Dijk M, et al. Continual learning strategies for cancer-independent detection of lymph node metastases. *Med Image Anal*. 2023;85:102755. doi:10.1016/J.MEDIA.2023.102755

Figure legends

Figure 1. Study cohorts and fast-track algorithm development principles

A. Principle of the AI Diagnostic Tool for Lymph Node Metastasis Detection in Colorectal Specimens. B. Study Cohorts. Comments: *This number does not include an additional 111 slides that were not manually annotated and were only used for hard negative mining to automatically extract often small, difficult regions to enrich the training dataset. #Total number of unique slides excluding those scanned twice by two different scanners. C. Fast-Track Development Framework. In brief, a pre-trained algorithm[5] from the same domain but for primary tumors was used to generate precise pre-annotations, which were corrected by human analysts. After the first training, the training dataset was enriched through difficult regions (hard-negative mining, slides with inflamed fatty tissue) and relevant regions from the primary tumor dataset. With the fast-track framework, the annotation process saves several months. For details, see Methods. Abbreviations: MPP – micron per pixel.

Figure 2. Principle of lymph node metastasis detection in whole slide images. Shown are two cases with mucinous and conventional morphology. The developed AI algorithm allows for precise segmentation of seven different tissue classes (the eighth class is slide background). The legend provides an explanation of the color coding for different classes.

Figure 3. Further examples of lymph node metastasis detection in whole slide images. The AI algorithm we developed allows for precise segmentation of seven different tissue classes (the eighth class is slide background). The legend provides an explanation for the color coding of the different classes. Comments: *-Naturally,

signet ring cell carcinomas represent the most challenging cases. In this case, indicated by (*) a tumor region was detected as tumor stroma. This was sometimes evident focally in the signet ring carcinoma cases, as presented here; however, no single slide was falsely negatively misclassified due to this effect.

Figure 4. Clinical validation of the AI algorithm using international multi-institutional cohort of cases.

A. Principles of Construction of the Test Cohort (more details in Figure 1B). B. Statistical Metrics of Algorithm Performance in Different Cohorts. Note that in 5 out of 7 cohorts, the algorithm achieved a sensitivity and NPV (Negative Predictive Value) of 1.000. C. Single Slide Level Confusion Matrices Involving Ground Truth (Tumor/Benign) and AI Algorithm Evaluation Results. Detailed analysis of false negative and false positive misclassifications is provided in Figures 5 and 6. D. Analysis of the inference speed of the algorithm. All metrics are provided per slide, showing very high speed of single slide analysis.

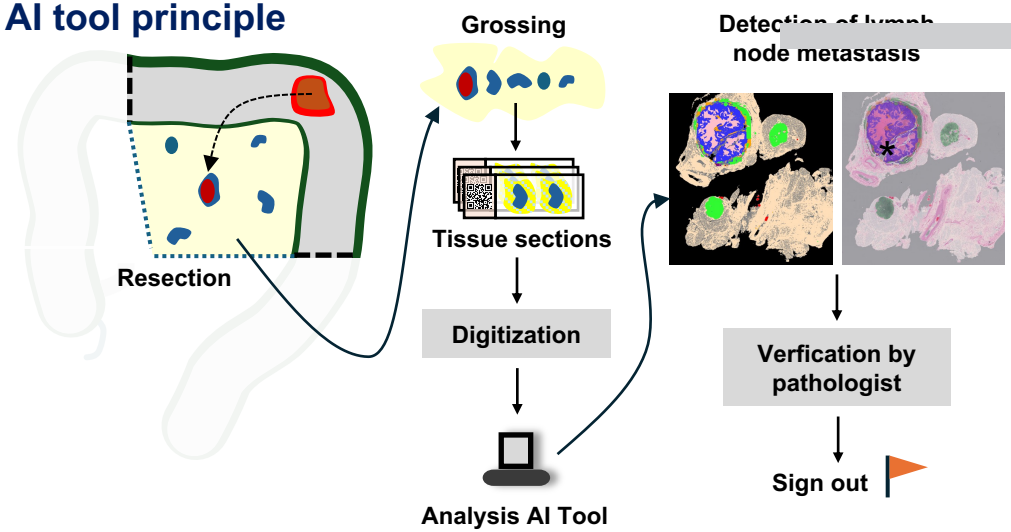
Figure 5. Analysis of false-negative misclassifications.

All regions with tumors that were formally classified as false negatives were correctly detected by the AI algorithm, but they fell below the initially selected, very small tumor area threshold. Three of the five misclassified slides contained only lymphangiosis, as shown in (C, D), while two contained a single gland invasive tumor component, as shown in (A, B). Cohort information is provided near each image with a brief description of the reason for misclassification.

Figure 6. Analysis of false positive misclassifications.

Most of the false positive misclassifications were considered very useful alerts (A-C) in cases of severely artificially altered tissue, suspicious fibrosis, or overactivated histiocytes, where immunohistochemistry might be necessary. A few misclassifications (D-F) were less useful and represented areas that might be included in the training dataset at a later stage to further improve the algorithm.

A AI tool principle



B Study cohorts

Cohort	Type	Slides, n	Cases, n	Slides with tumor, n	Histoscanner	Magnification	Scanner MPP
UKK A	Train	432*		111	Hamamatsu S360	40x	0.2305
WNS A	Test	795	46	90	Hamamatsu S360	40x	0.2305
WNS B	Test	789	46	93	Leica GT450	40x	0.2628
CHA	Test	1587	135	197	Leica GT450	40x	0.2628
UKK B	Test	4911	207	492	Hamamatsu S360	40x	0.2305
UKK C	Test	4704	207	428	Leica GT450	40x	0.2628
KAM	Test	652	59	80	Philips UltraFast	40x	0.5000
ESS	Test	1022	79	103	3D HISTECH	40x	0.2426
Total Test#:	Test	8967	526	965			

C Fast-track development principle

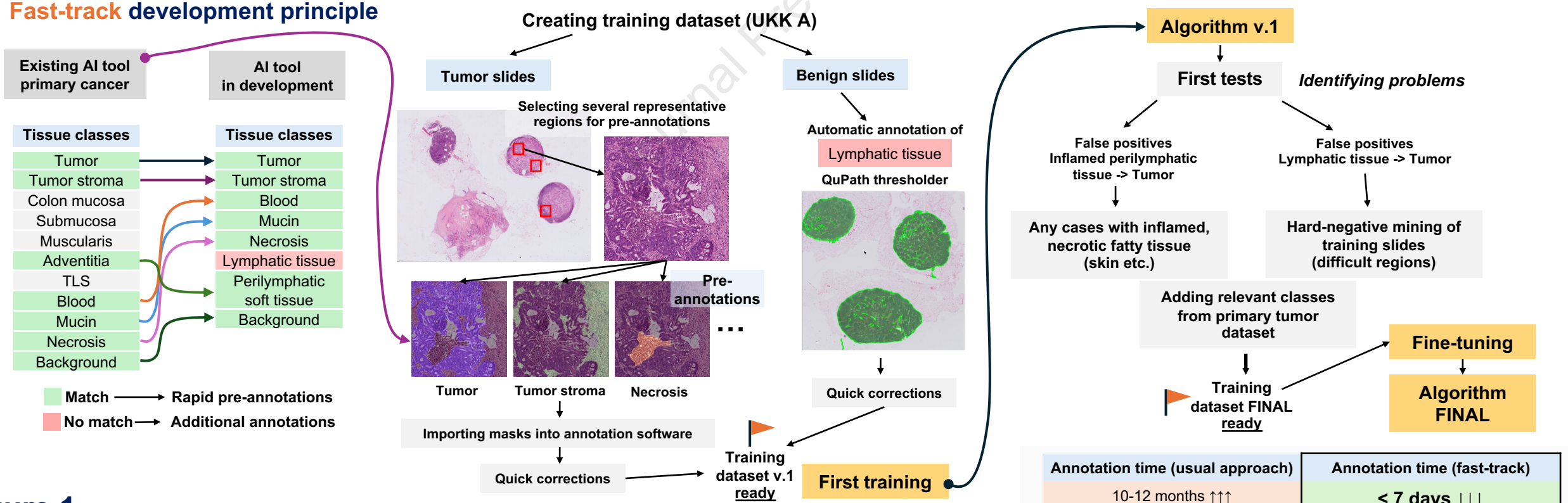


Figure 1

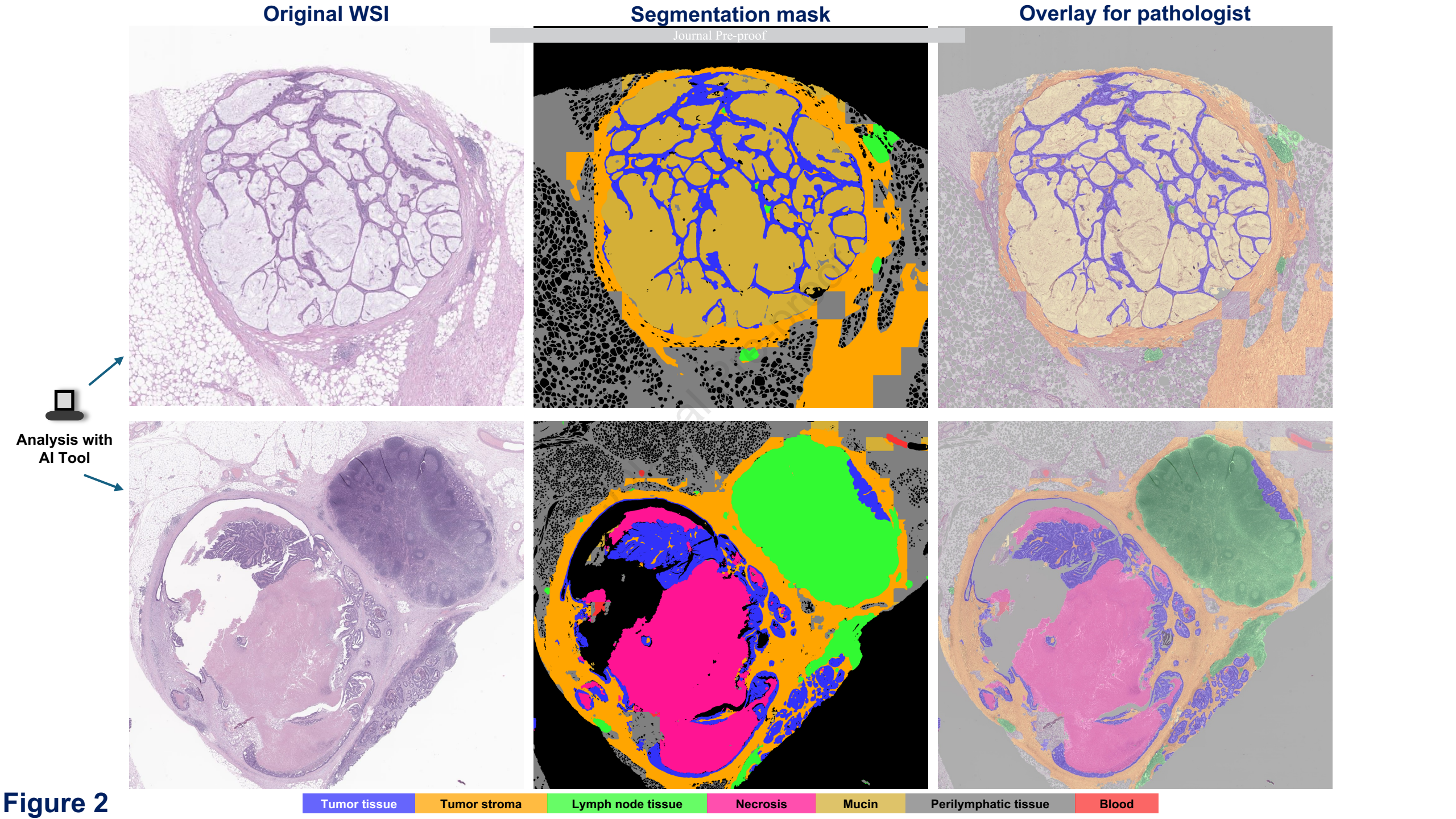


Figure 2

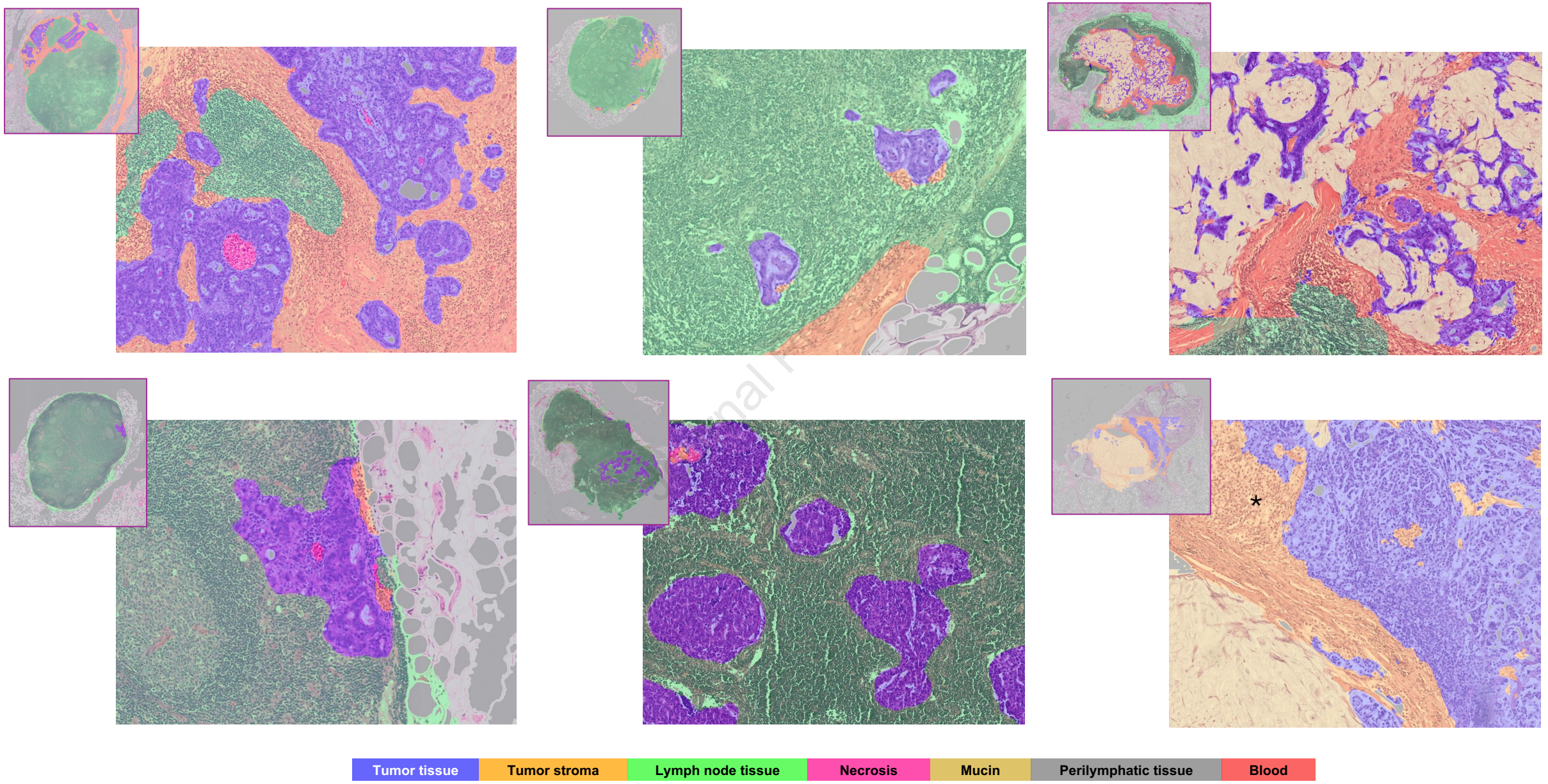
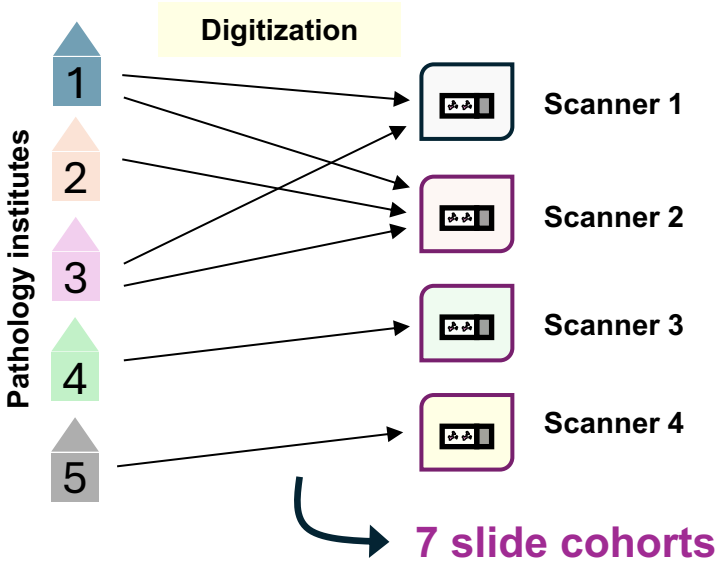


Figure 3

A Multi-institutional validation



B Clinical metrics (slide level)

Cohort	ACC	F1	PPV	NPV	SENS	SPEC
WNS A	0.923	0.747	0.596	1.000	**1.000**	0.913
WNS B	0.987	0.949	0.903	1.000	**1.000**	0.986
CHA	0.983	0.935	0.894	0.997	**0.980**	0.983
UKK B	0.978	0.899	0.817	1.000	**1.000**	0.975
UKK C	0.991	0.953	0.911	1.000	**1.000**	0.990
KAM	0.949	0.827	0.712	0.998	**0.988**	0.944
ESS	0.970	0.869	0.769	1.000	**1.000**	0.966

ACC accuracy **F1** F1 score **PPV** positive predictive value
NPV negative predictive value **SENS** sensitivity **SPEC** specificity

C Analysis of single slides

WNS A	AI: Ben	AI: Tu
GT: Ben	644	61
GT: Tu	0	90

UKK B	AI: Ben	AI: Tu
GT: Ben	4309	110
GT: Tu	0	492

CHA	AI: Ben	AI: Tu
GT: Ben	1367	23
GT: Tu	4	193

WNS B	AI: Ben	AI: Tu
GT: Ben	686	10
GT: Tu	0	93

UKK C	AI: Ben	AI: Tu
GT: Ben	4234	42
GT: Tu	0	428

KAM	AI: Ben	AI: Tu
GT: Ben	540	32
GT: Tu	1	79

ESS	AI: Ben	AI: Tu
GT: Ben	888	31
GT: Tu	0	103

GT ground truth
AI AI tool
Tu Slide without tumor
Ben Slide with tumor

D Time-to-analysis metrics (time per slide)

Cohort	Min	Max	Median	Mean
WNS A	<5 s	306 s	30 s	50 s
WNS B	<5 s	276 s	30 s	43 s
CHA	<5 s	276 s	30 s	50 s
UKK B	<5 s	324 s	24 s	32 s
UKK C	<5 s	282 s	18 s	23 s
KAM	<5 s	354 s	24 s	35 s
ESS	<5 s	396 s	30 s	49 s

Figure 4

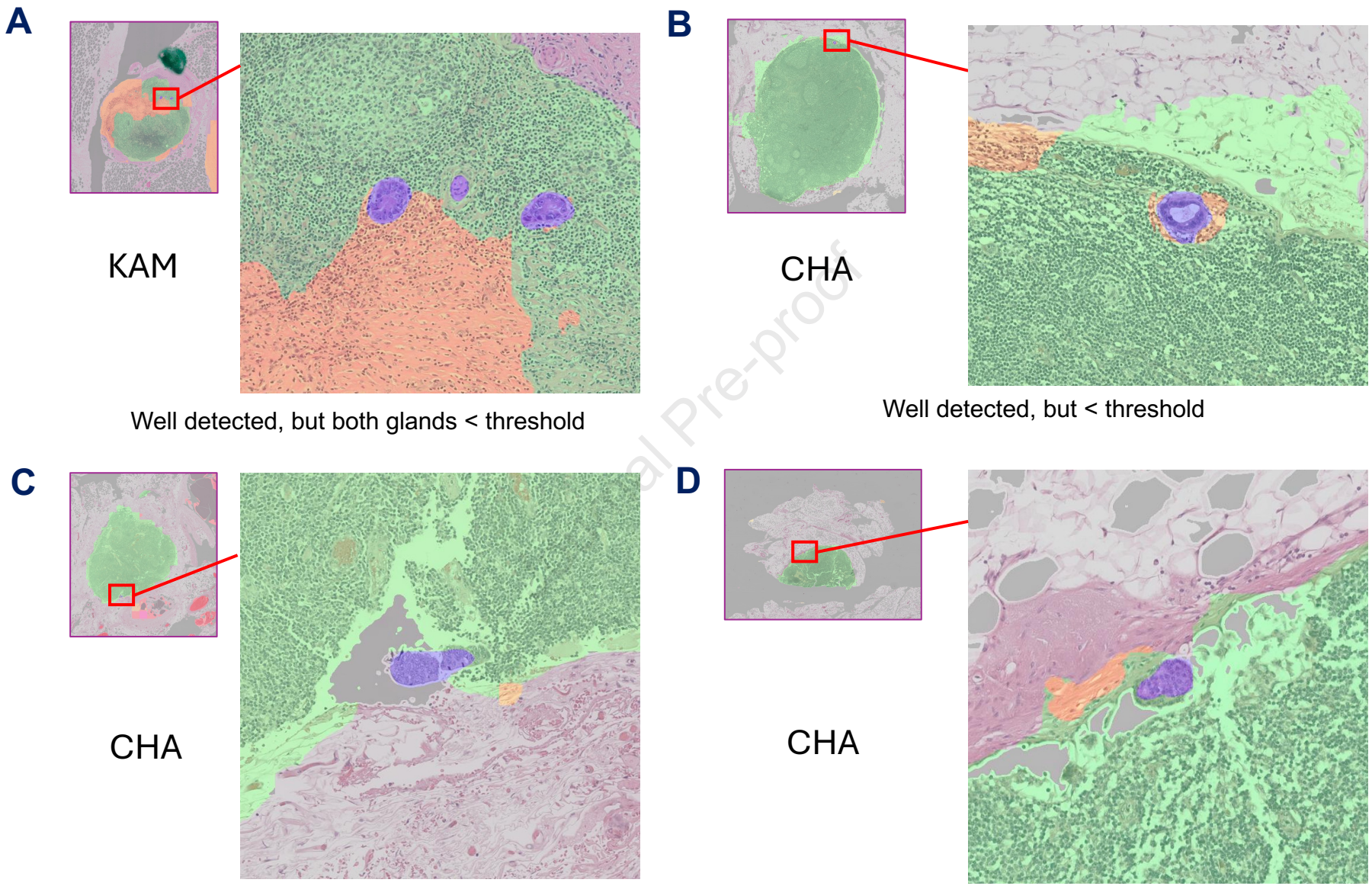
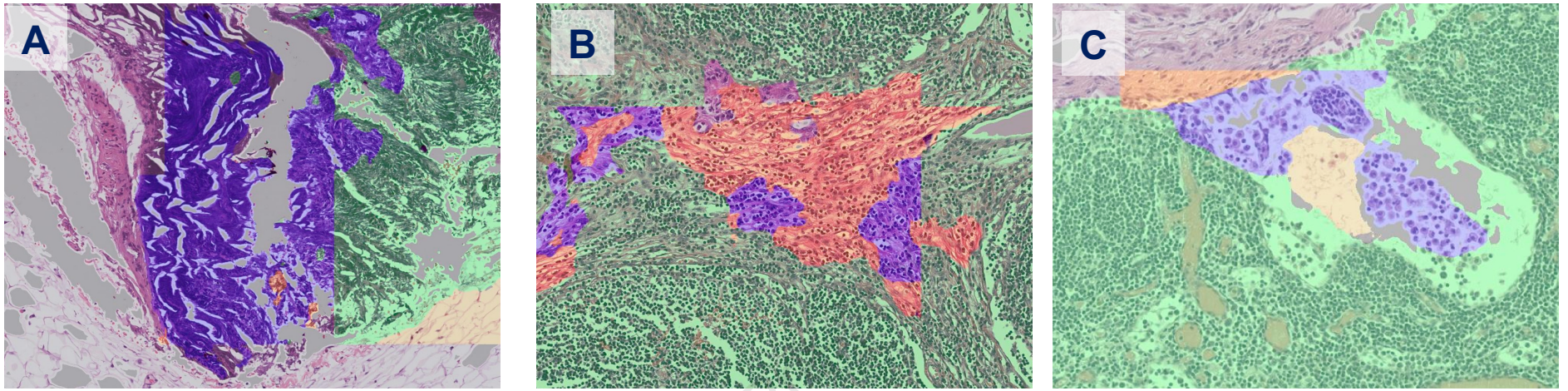


Figure 5

Useful false positive alerts

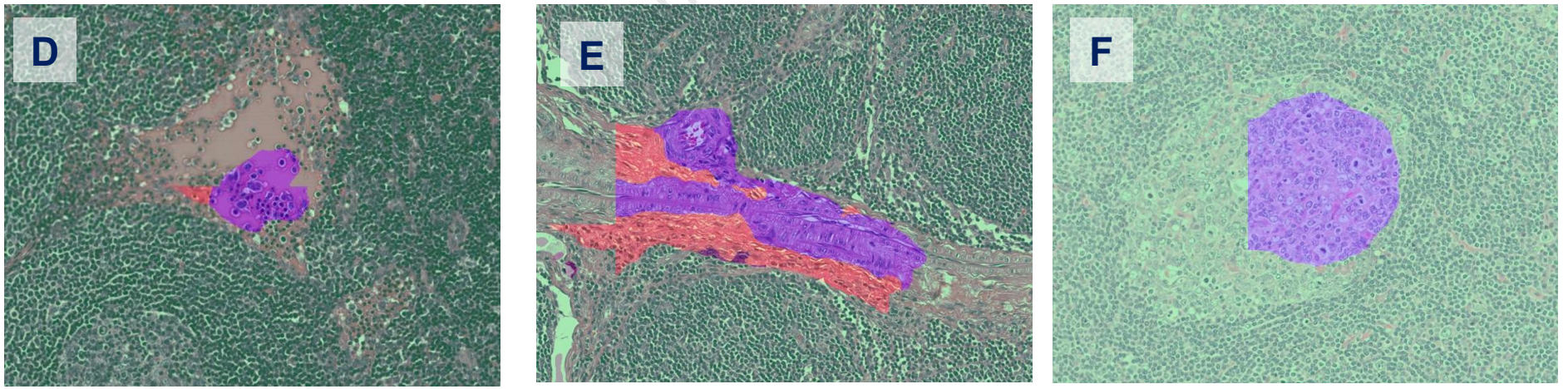


Severe artifacts

Suspicious fibrosis with inflammation

Overactivated histiocytes

Less useful false positive classifications



Blood vessels, histiocytes

Blood vessels

Activated germ centers

Figure 6