Journal of Pathology
J Pathol 2021; **00:** 000–000
Published online 22 October 2021 in Wiley Online Library (wileyonlinelibrary.com) **DOI:** 10.1002/path.5800

ORIGINAL PAPER

Weakly supervised annotation-free cancer detection and prediction of genotype in routine histopathology

Peter Leonard Schrammen[1], Narmin Ghaffari Laleh[1], Amelie Echle[1], Daniel Truhn[2], Volkmar Schulz[3,4,5,6], Titus J Brinker[7], Hermann Brenner[8,9,10], Jenny Chang-Claude[11,12], Elizabeth Alwers[8], Alexander Brobeil[13,14], Matthias Kloor[13], Lara R Heij[15,16,17], Dirk Jäger[18], Christian Trautwein[1], Heike I Grabsch[15,19], Philip Quirke[19], Nicholas P West[19], Michael Hoffmeister[8] and Jakob Nikolas Kather[1,10,18,19]*

[1] *Department of Medicine III, University Hospital RWTH Aachen, Aachen, Germany*
[2] *Department of Radiology, University Hospital RWTH Aachen, Aachen, Germany*
[3] *Department of Physics of Molecular Imaging Systems, Experimental Molecular Imaging, RWTH Aachen University, Aachen, Germany*
[4] *Fraunhofer Institute for Digital Medicine MEVIS, Bremen, Germany*
[5] *Comprehensive Diagnostic Center Aachen (CDCA), University Hospital Aachen, Aachen, Germany*
[6] *Hyperion Hybrid Imaging Systems GmbH, Aachen, Germany*
[7] *Digital Biomarkers for Oncology Group, German Cancer Research Center (DKFZ), Heidelberg, Germany*
[8] *Division of Clinical Epidemiology and Aging Research, German Cancer Research Center (DKFZ), Heidelberg, Germany*
[9] *Division of Preventive Oncology, German Cancer Research Center (DKFZ) and National Center for Tumor Diseases (NCT), Heidelberg, Germany*
[10] *German Cancer Consortium (DKTK), German Cancer Research Center (DKFZ), Heidelberg, Germany*
[11] *Division of Cancer Epidemiology, German Cancer Research Center (DKFZ), Heidelberg, Germany*
[12] *Cancer Epidemiology Group, University Cancer Center Hamburg, University Medical Center Hamburg-Eppendorf, Hamburg, Germany*
[13] *Institute of Pathology, University Hospital Heidelberg, Heidelberg, Germany*
[14] *Tumor Bank Unit, Tissue Bank of the National Center for Tumor Diseases, Heidelberg, Germany*
[15] *Department of Pathology, GROW School for Oncology and Developmental Biology, Maastricht University Medical Center+, Maastricht, The Netherlands*
[16] *Institute of Pathology, University Hospital RWTH Aachen, Aachen, Germany*
[17] *Department of Surgery and Transplantation, University Hospital RWTH Aachen, Aachen, Germany*
[18] *Medical Oncology, National Center of Tumor Diseases (NCT), University Hospital Heidelberg, Heidelberg, Germany*
[19] *Pathology & Data Analytics, Leeds Institute of Medical Research at St James's, University of Leeds, Leeds, UK*

**Correspondence to: JN Kather, Department of Medicine III, University Hospital Aachen, Pauwelsstraße 30, 52074 Aachen, Germany. E-mail: jkather@ukaachen.de*

Abstract

Deep learning is a powerful tool in computational pathology: it can be used for tumor detection and for predicting genetic alterations based on histopathology images alone. Conventionally, tumor detection and prediction of genetic alterations are two separate workflows. Newer methods have combined them, but require complex, manually engineered computational pipelines, restricting reproducibility and robustness. To address these issues, we present a new method for simultaneous tumor detection and prediction of genetic alterations: The Slide-Level Assessment Model (SLAM) uses a single off-the-shelf neural network to predict molecular alterations directly from routine pathology slides without any manual annotations, improving upon previous methods by automatically excluding normal and non-informative tissue regions. SLAM requires only standard programming libraries and is conceptually simpler than previous approaches. We have extensively validated SLAM for clinically relevant tasks using two large multicentric cohorts of colorectal cancer patients, Darmkrebs: Chancen der Verhütung durch Screening (DACHS) from Germany and Yorkshire Cancer Research Bowel Cancer Improvement Programme (YCR-BCIP) from the UK. We show that SLAM yields reliable slide-level classification of tumor presence with an area under the receiver operating curve (AUROC) of 0.980 (confidence interval 0.975, 0.984; $n = 2{,}297$ tumor and $n = 1{,}281$ normal slides). In addition, SLAM can detect microsatellite instability (MSI)/mismatch repair deficiency (dMMR) or microsatellite stability/mismatch repair proficiency with an AUROC of 0.909 (0.888, 0.929; $n = 2{,}039$ patients) and *BRAF* mutational status with an AUROC of 0.821 (0.786, 0.852; $n = 2{,}075$ patients). The improvement with respect to previous methods was validated in a large external testing cohort in which MSI/dMMR status was detected with an AUROC of 0.900 (0.864, 0.931; $n = 805$ patients). In addition, SLAM provides human-interpretable visualization maps, enabling the analysis of multiplexed network predictions by human experts. In summary, SLAM is a new simple and powerful method for computational pathology that could be applied to multiple disease contexts.

© 2021 The Authors. *The Journal of Pathology* published by John Wiley & Sons, Ltd. on behalf of The Pathological Society of Great Britain and Ireland.

Keywords: artificial intelligence; deep learning; colorectal cancer; computational pathology; digital pathology; microsatellite instability; Lynch syndrome

Received 12 May 2021; Revised 1 August 2021; Accepted 3 September 2021

© 2021 The Authors. *The Journal of Pathology* published by John Wiley & Sons, Ltd. on behalf of The Pathological Society of Great Britain and Ireland. This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

Conflict of interest statement: JNK declares consulting services for Owkin, France and Panakeia, UK. TJB reports owning a company that develops mobile apps, outside the scope of the submitted work (Smart Health Heidelberg GmbH). PQ has had paid roles in the English NHS bowel cancer screening program over the course of this study. No other potential conflicts of interest were declared.

Introduction

Colorectal cancer (CRC) is one of the most common types of cancer and one of the top causes of cancer mortality [1]. In routine clinical workflows, CRC is diagnosed by histopathologic evaluation of H&E-stained tissue slides. In addition, all patients with metastatic or unresectable CRC are recommended to undergo testing for microsatellite instability (MSI) or mismatch repair deficiency (dMMR) and should be tested for mutations of the *KRAS* and *BRAF* genes [2]. In the UK, all patients with CRC, irrespective of tumor stage, are recommended to undergo MSI or dMMR testing [3]. Metastatic MSI/dMMR CRC are directly targetable by cancer immunotherapy, which is currently approved as a first-line therapeutic approach to this disease subtype [4]. MSI, as determined by polymerase chain reaction (PCR), and dMMR, as determined by immunohistochemistry (IHC), are used interchangeably in most clinical situations, although the results of these different tests are not always concordant [3,5]. Another type of clinically relevant genetic alteration in CRC is a mutated *BRAF* gene in metastatic CRC, which is directly targetable in a second-line therapeutic setting [6]. Currently, diagnosis of cancer on histopathology images and genetic testing on tumor tissue form two distinct laboratory workflows: although they are both coordinated by the pathologist as a central coordinator, they are performed using different laboratory methods. However, increasing efforts to digitize routine histopathology workflows [7,8] will potentially make digitized whole-slide images (WSI) routinely available in the future. Recent studies have shown that a wide range of molecular features, including MSI/dMMR status and *BRAF* mutational status, can be predicted from digitized slides of CRC using deep learning, an artificial intelligence technology [9–14]. The application of such methods is not limited to CRC but has been demonstrated in bladder cancer [15], breast cancer [16,17], sarcoma [18], head and neck cancer [19], hepatocellular carcinoma [20], and several other types of solid tumor [8,12]. Therefore, in the future, deep learning could supplement current molecular testing strategies in solid tumors and could be used as a tool for translational research [21].

Multiple different technical pipelines have been proposed to infer molecular alterations from WSI and each of them has limitations [8]. The first scientific publications in deep learning-based molecular subtyping in 2018 and 2019 applied a simple tumor annotation-based 'majority vote', i.e. they were based on a two-step process: first, they located tumor tissue in the tissue section based on manual [22] or automatic segmentations and, subsequently, the tumor tissue was processed by another neural network [10]. Further studies showed that a manual annotation-based approach could yield very high performance for tumor detection on large datasets [23]. More recent studies used deep learning to predict genotypes directly from the whole slide, including tumor and non-tumor tissue. These so-called weakly supervised approaches do not require any explicit tumor detection, applying a simple whole-slide majority vote [9,12]. Such approaches have achieved a high performance for the prediction of molecular alterations, but they sacrifice interpretability. Using the whole tissue to predict molecular features in the tumor tissue imposes predictions of molecular changes on non-tumor regions such as normal mucosa, which may not be useful or desirable. More recent studies have addressed this issue by using a new technology based on multiple-instance learning: in the context of prostate cancer detection, a weakly supervised approach yielded a clinical grade performance [24] and is currently being marketed as a commercial product [25]. Other attention-based approaches were recently applied to tumor detection in various cancer types. For example, clustering-constrained attention multiple instance learning has been proposed as a powerful methods pipeline for tumor detection and determining histopathologic subtypes [26,27]. However, attention-based multiple learning is not widely used for predicting molecular alterations from image data, possibly because these models are complex and data-hungry [24].

A general observation is that methods pipelines in computational pathology become more and more intricate: they require hand-crafted network models, loss functions, and intricate pre-/post-processing pipelines, which cannot be easily implemented using standard programming libraries [24,28–30]. In particular, the custom architectures and data loader required for these methods are not available out-of-the-box in popular machine learning environments, such as PyTorch, TensorFlow, Keras, or Fastai. This is in stark contrast to initial publications, which were easily re-implementable in standard programming environments in a few lines of code [10,22]. As complex workflows limit widespread reproduction and adoption, there is a need for powerful, adaptable, easily implementable, end-to-end methods for molecular testing of cancer.

Therefore, in this study, we sought to combine the ease-of-use of off-the-shelf models with one-stop-shop convenience and improved interpretability. At the same time, we strived to deliver the first application of deep learning for one-shot tumor localization and genetic subtyping in CRC. In other words, we aimed to unify the workflows of tumor diagnosis and subtyping in a single-pass neural network.

Materials and methods

Ethics statement and patient cohorts

For this study we used anonymized H&E-stained slides of colorectal adenocarcinoma of two large cohorts.

© 2021 The Authors. *The Journal of Pathology* published by John Wiley & Sons, Ltd. on behalf of The Pathological Society of Great Britain and Ireland. www.pathsoc.org

To train the neural network we used digitized tumor-bearing tissue slides from the Darmkrebs: Chancen der Verhütung durch Screening (DACHS) study ($n = 2{,}448$ patients), a large population-based case–control and patient cohort study on CRC, including samples of patients with stages I–IV from different laboratories in southwestern Germany. We received and used exactly one tumor-bearing tissue slide per patient. For $n = 1{,}281$ of these patients, an additional non-tumor slide was available, i.e. a tissue slide extracted from the same surgical specimen but containing only normal colon mucosa, submucosa, and smooth muscle tissue. This 'normal' tissue slide was used as an additional input for the deep learning model, as explained below. Use of the DACHS tissue samples for scientific purposes was approved by the ethics committees of Heidelberg University and the medical boards of Rhineland-Palatinate and Baden-Württemberg, with the written informed consent of all participants [31]. The digitized tissue slides were provided by the Tissue Bank of the National Center for Tumor Diseases (Heidelberg, Germany) in accordance with the regulations of the tissue bank. Some tissue slides in the DACHS cohort had blue and/or black pen marks circling tumor tissue and/or normal tissue on the slide. MSI status in the DACHS cohort was investigated using a three-plex PCR panel, as described previously [32]. For external validation we applied the deep learning system on H&E-stained slides derived from the population-based Yorkshire Cancer Research Bowel Cancer Improvement Programme (YCR-BCIP) [33], comprising 889 patients who had surgical resection. dMMR or mismatch repair proficiency (pMMR) was determined with a standard four-plex IHC assay on whole slides. No pen marks were present on the slides in the YCR-BCIP dataset. The clinicopathologic characteristics of all patients are summarized in Table 1. Glass slides in DACHS and YCR-BCIP were digitized with Leica Aperio scanners (Leica Biosystems, Wetzlar, Germany) using a 20× objective and were saved as SVS files with JPEG compression. We received and used exactly one digitized tumor slide from each patient in the DACHS cohort and the YCR-BCIP cohorts. Only patients with an available H&E slide and clinicopathologic features were used for the analysis. Some samples were excluded due to missing clinicopathologic data or missing WSI. Sample flowcharts for all experiments are provided in supplementary material, Figure S1.

Image preprocessing pipeline

Non-overlapping image tiles with a size of 512 × 512 pixels with a resolution of 0.5 µm per pixel were extracted from the WSIs. Tiles with background (more than 50% white area on the tile), blurry artifacts, and pen marks were removed during the tessellation process. The standard deviation of each color channel in a tile and the average detected edges using canny edge detection of OpenCV package in Python 3.8 were used to detect these tiles. To remove the bias of different staining procedures, all tiles were normalized based on one reference image using the Macenko normalization method using a reference image that is publicly available at: https://raw.githubusercontent.com/jnkather/DeepHistology/master/subroutines_normalization/Ref.png [34]. After this step, tiles were used as an input for the neural network. Whenever a slide contributed more than 1,000 tiles, only 1,000 randomly chosen tiles were used. The source code for data preprocessing is available under an open-source license at: https://github.com/KatherLab/preProcessing. For all experiments, only patient-level labels were used and all tiles in the training sets were assumed to inherit the label of their parent patient. To mitigate class imbalance in the patient labels during training, tiles from the more abundant class were randomly undersampled. This means that for training neural networks, equal numbers of tiles from the positive and negative classes were used and classifiers were trained on tile-level-balanced image sets. For deployment of classifiers to the test partition in cross-validation or to the external validation set, no such class balancing procedure was applied.

Algorithm

Here, we propose a new method, the Slide-Level Assessment Model (SLAM). We assume that colorectal tumors can carry a feature of interest, the 'target', which is defined on the level of patients. The aim is to determine the presence of the target directly from a digitized glass slide (WSI). In the present study we explored the following targets: *BRAF* status (mutated or non-mutated), MSI/MMR status (MSI/dMMR or microsatellite stability [MSS]/pMMR), and grade of differentiation (high grade, comprising poorly differentiated and undifferentiated [grade 3–4] and low grade, comprising well and moderately differentiated [grade 1–2]). For all targets, only slide-level labels, not tile labels, are available. We assume that only the tumor tissue carries information related to these labels, but tumor-bearing slides usually contain some non-tumor tissue adjacent to the tumor. The state of the art (SOTA) model is to train end-to-end deep learning systems on all tiles generated from these WSIs, tumor and non-tumor [9,12]. This is potentially suboptimal as it dilutes the information of interest and assigns a prediction score for non-tumor tiles. Although some studies solve this problem with manual annotations [23] or adding a separate network for tumor detection [10], SLAM solves this in a single step: SLAM uses an end-to-end neural network based on ShuffleNet, a lightweight off-the-shelf model [35]. The output layer has been modified to have three output classes: tumor tissue belonging to the positive class (mutated, MSI/dMMR, high grade, etc.), tumor tissue belonging to the negative class (non-mutated, MSS/pMMR, low grade, etc.), and non-tumor tissue, which is assumed to be non-informative regarding the presence of the target class (label). This procedure can be extended to an arbitrary number of target classes. We used WSI with slide-level labels for training. Based on the ground truth

© 2021 The Authors. *The Journal of Pathology* published by John Wiley & Sons, Ltd. on behalf of The Pathological Society of Great Britain and Ireland. www.pathsoc.org

Table 1. Clinicopathologic features of the patient cohorts. Only samples with matched histopathology images and clinical data were included. For all cohorts and all targets, sample flowcharts are available in supplementary material, Figure S1.

	DACHS tumor	YCR-BCIP
Number of patients	2,448	889
Region	Southwest Germany	Yorkshire, UK
MSI or dMMR	PCR 3-plex	IHC 4-plex
Male	1,436 (58.7%)	494 (55.6%)
Female	1,012 (41.3%)	395 (44.4%)
Gender data unavailable	0 (0.0%)	0 (0.0%)
Gender data available	2,448 (100%)	889 (100%)
Samples included	2,297 (93.8% of 2,448)	-
Males in included patients	1,345	-
Females in included patients	952	-
Age < 40 years	24 (1.0%)	-
Age 40–50 years	99 (4.0%)	-
Age 50–60 years	359 (14.7%)	157 (17.7%)
Age 60–70 years	780 (31.9%)	241 (27.1%)
Age 70–80 years	801 (32.7%)	316 (35.5%)
Age 80+ years	385 (15.7%)	175 (19.7%
Colon cancer	1,488 (60.8%)	669 (75.3%)
Rectal cancer	960 (39.2%)	216 (24.3%)
Stage I	485 (19.8%)	169 (19.0%)
Stage II	801 (32.7%)	317 (35.7%)
Stage III	822 (33.6%)	370 (42.6%)
Stage IV	337 (13.8%)	0 (0.0%)
Adjuvant chemotherapy	1,043 (42.6%)	Unknown
No adjuvant chemotherapy	1,389 (56.7%)	Unknown
Adjuvant radiotherapy	2,250 (91.9%)	Unknown
After recurrence	2 (0.1%)	Unknown
No adjuvant radiotherapy	187 (7.6%)	Unknown
Neoadjuvant therapy	281 (11.5%)	Unknown
No neoadjuvant therapy	2,159 (88.2%)	Unknown
Low grade (grade 1–2)	1,587 (64.8%)	Unknown
High grade (grade 3–4)	561 (22.9%)	Unknown
Grade unavailable	300 (12.3%)	-
Grade available	2,148 (87.7%)	-
Samples included	2,066 (84.4% of 2,448)	-
Low grade in included patients	1,518	-
High grade in included patients	548	-
BRAF mutation	151 (6.2%)	75 (8.4%)
BRAF wild type	1,930 (78.8%)	32 (3.6%)
BRAF status unavailable	367 (14.9%)	782 (88.0%)
BRAF available	2,081 (85.0%)	107 (12.0%)
Samples included	2,075 (84.8% of 2,448)	-
BRAF mutation in included patients	151	-
BRAF wild type in included patients	1,924	-
KRAS mutation	677 (27.6%)	Unknown
KRAS wild type	1,397 (57.1%)	Unknown
KRAS status unavailable	374 (15.3%)	-
KRAS status available	2,074 (84.7%)	-
Samples included	2,068 (84.5% of 2,448)	-
KRAS mutation in included patients	674	-
KRAS wild type in included patients	1,394	-
MSI/dMMR	210 (8.6%)	117 (14.39%)
MSS/pMMR	1,836 (75.0%)	772 (86.8%)
MSI/dMMR status unavailable	402 (16.4%)	0 (0.0%)
MSI/dMMR status available	2,046 (83.5%)	889 (100.0%)
Samples included	2,039 (83.3% of 2,448)	805 (90.6% of 889)
MSI/dMMR in included patients	210	112
MSS/pMMR in included patients	1,829	693

labels, each slide was assigned to one of these three classes: 'positive' tumor slides (containing tumor tissue in the positive class as well as some non-tumor/non-informative tissue), 'negative' tumor slides (containing tumor tissue in the negative class as well as some non-tumor/non-informative tissue), and non-tumor slides (containing only non-tumor/non-informative tissue). All image tiles generated from the slides inherited the

© 2021 The Authors. *The Journal of Pathology* published by John Wiley & Sons, Ltd. on behalf of The Pathological Society of Great Britain and Ireland. www.pathsoc.org

slide-level label (positive, negative, or non-tumor/non-informative) and were used to train the network. Thus, even though the training sets were contaminated with non-tumor/non-informative tissue, the SLAM network can learn to distinguish tumor tissue from non-tumor/non-informative tissue because the non-tumor tissue is introduced as an explicit third class. When deployed to an image in the test set, each tile is assigned a probability value (tile-level soft prediction). The class with the highest probability value for each tile is used for all further steps (tile-level hard prediction). Thus, each tile is assigned a single prediction category by the deep learning SOTA (positive or negative) model or SLAM (positive, negative, or non-informative). SOTA methods have also been applied to multiclass problems [9], in which the performance for each target class is obtained by a one-versus-rest procedure. Like SOTA, SLAM is able to handle such multiclass problems. For simplicity, we only refer to the (much more common) binary classification problem from now on. For N_pos being the number of mutated tiles and N_tot being the total number of tiles, the patient prediction scores (PPS) in SOTA [9] are defined as follows: PPS = N_pos/N_tot. However, it is known that N_tot is contaminated by non-informative tiles corresponding to normal tissue. Therefore, N_tot is artificially inflated if there is a relevant amount of non-tumor tissue on the slide. This is solved by SLAM, which predicts positive tumor tiles (N_pos), negative tumor tiles (N_neg), and non-tumor or non-informative tiles (N_nt). PPS in SLAM are calculated as PPS = N_pos/(N_tot − N_nt). Technical details are listed in supplementary material, Table S1 and an additional description of SLAM is provided in supplementary material, Figure S2. In this study, we compared the performance of SLAM to the SOTA algorithm.

Experimental design and statistics

First, we tested whether tumor slides and non-tumor slides could be distinguished with a high accuracy. Then, we trained SLAM on five binary classification tasks in a within-cohort approach by using patient-level three-fold cross-validation in the DACHS cohort. The classification targets were grade (low/high), gender (female/male), *KRAS* mutation (mutated/wild type), *BRAF* status (mutated/wild type), and MSI/MMR status (MSI/MSS or dMMR/pMMR). Gender was included as a negative control. Although MSI and dMMR are measured by different laboratory methods (PCR and IHC, respectively) and are not 100% overlapping, they are widely regarded as synonymous for clinical decision making. Therefore, here we refer to 'MSI/dMMR status'. Finally, we validated the model trained on DACHS on an external cohort, YCR-BCIP, for prediction of MSI/dMMR status. The primary statistical endpoint was the area under the receiver operating curve (AUROC) with 100-fold bootstrapped confidence intervals. This means that the confidence intervals were obtained by a procedure in which a list of PPSs was generated 100 times, AUROCs were re-calculated, and the 95% confidence interval on this distribution is given. Each time the list of prediction scores was generated, n patients were randomly chosen from the list of N patients with replacement. This procedure was performed by the Matlab function 'perfcurve', which is documented at https://www.mathworks.com/help/stats/perfcurve.html. Secondary statistical endpoints were accuracy, sensitivity, specificity, and F1 score of the SOTA model and SLAM. To generate a cut-off value for these statistics, an identical automatic procedure was applied to the patient-level prediction scores in each experiment. Using the ROC curve, the closest threshold value corresponding to a sensitivity of 80% was identified and rounded to three decimal places. Subsequently, using this threshold value, a confusion matrix and statistics were calculated. Because ROC curves are not continuously defined, the final sensitivity could differ from 80% (see supplementary material, Table S2).

Visualization

To visualize three classes in a single visualization, we employed multiplexed heat maps using three base color vectors to achieve close to perceptually optimized color maps, as described previously [36]. Based on each tile prediction value z for MSI (z_{MSI}), MSS (z_{MSS}), and normal (z_{normal}), the color C was generated with three red, green, blue (RGB) color vectors c ($c_{MSI} = [0.8,0,0]$, $c_{MSS} = [1,1,0]$, $c_{normal} = [0,0,1]$) as follows: $C = z_{MSI} * c_{MSI} + z_{MSS} + c_{MSS} + z_{normal} * c_{normal}$. To generate smooth maps from sparse predictions, we interpolated between the z values on a regular two-dimensional grid. In addition, we selected the highest scoring tiles (based on tile-level soft predictions) for the highest-scoring patients (based on patient predictions) and reviewed these tiles with a pathologist to identify human-interpretable morphologic patterns of interest.

Results

Automatic slide-level tumor detection and grading

Here we present SLAM (Figure 1A–C). First, we assessed the ability of SLAM to automatically detect tumor-bearing slides on a slide level using weak (slide-level) labels with a three-fold cross-validation approach, using WSIs containing both tumor and non-tumor tissue as well as normal tissue (non-tumor colorectal tissue) slide images without any tumor tissue. In DACHS (n = 2,448 patients, Table 1), this achieved a high slide-level AUROC of 0.980 (0.975, 0.984). This demonstrates that tumor-bearing and non-tumor-bearing tissue slides can be well distinguished by a neural network trained on slide-level labels. Next, we moved from two classes to three classes and explored SLAM's accuracy in tumor grading. Ground truth labels were the predominant histopathologic grades of differentiation in each slide measured according to local standard procedures, binarized into low grade (grades 1–2) and

© 2021 The Authors. *The Journal of Pathology* published by John Wiley & Sons, Ltd. on behalf of The Pathological Society of Great Britain and Ireland. www.pathsoc.org

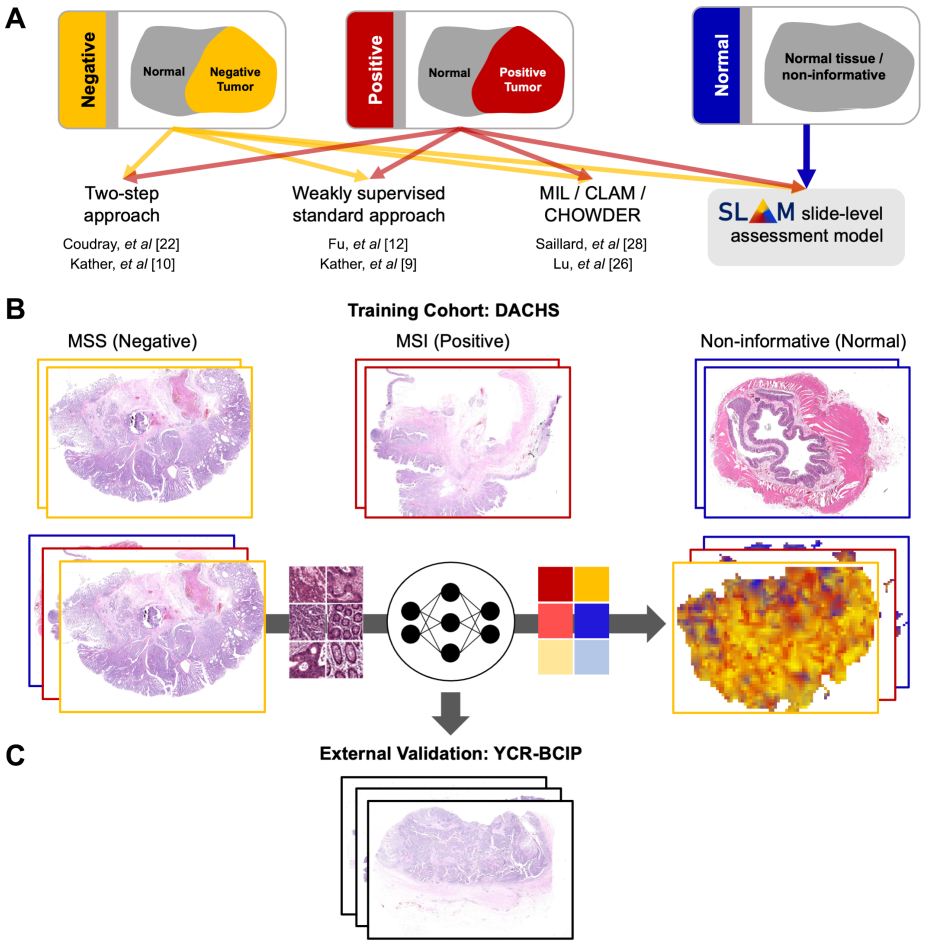

Figure 1. SLAM. (A) Previous studies used positive and negative tumors with slide-level labels, whereas SLAM includes normal tissue slides in the training. (B) Histologic routine images of tumor (yellow, MSS; red, MSI) and normal tissue (H&E) from DACHS ($n = 2{,}039$ for MSI status) were collected, tessellated, and normalized. SLAM was used to provide patient-level predictions and multiplexed spatial prediction maps. (C) External validation of SLAM on YCR-BCIP ($N = 805$ for MSI status).

high grade (grades 3–4). Applying the SOTA model on the grading target resulted in an AUROC of 0.722 (0.699, 0.744; see supplementary material, Figure S3A), which was improved by SLAM to an AUROC of 0.751 (0.727, 0.774; see supplementary material, Table S2). As a negative control, i.e. a target feature that should not be detectable, we included gender. Indeed, neither the baseline model nor SLAM could reliably infer gender from raw histopathology slides, reflected by AUROCs close to 0.50 (see supplementary material, Table S2).

SLAM outperforms classical deep learning models for prediction of genetic alterations

Next, we applied SLAM to the detection of two clinically targetable molecular alterations in CRC: MSI/dMMR status and *BRAF* mutational status. For MSI/dMMR detection, SLAM achieved an AUROC of 0.909 (0.888, 0.929, Figure 2A, supplementary material, Table S2), being more accurate than the deep learning-based SOTA model with an AUROC of 0.879 (0.855, 0.902). Compared with the previous deep learning SOTA, SLAM increased the F1 score from 0.477 to 0.559 and also increased the accuracy, sensitivity, and specificity of the predictions (see supplementary material, Table S2 and confusion matrices in supplementary material, Figure S4). Similarly, for inference of *BRAF* mutational status based on slide-level labels, SOTA achieved an AUROC of 0.782 (0.736, 0.813), which was improved to 0.821 (0.786, 0.852) by SLAM (see supplementary material, Figure S3B and Table S2). Again, accuracy, sensitivity, specificity and F1 score were also improved. Taken together, these data show that SLAM improves detection performance of molecular subtypes compared with SOTA. Importantly, we found that performance for prediction of MSI/dMMR status and *BRAF* mutational status particularly increased from SOTA to SLAM in the high-sensitivity region of the classification model, i.e. the upper region of the ROC curve (Figure 2A and supplementary material, Figure S3B). In addition, we evaluated whether *KRAS* mutational status was predictable from tissue slides. Previous studies have shown only a low predictability of *KRAS* status from slides by previous approaches [9], which in our experiments was reflected by a poor

© 2021 The Authors. *The Journal of Pathology* published by John Wiley & Sons, Ltd. on behalf of The Pathological Society of Great Britain and Ireland. www.pathsoc.org

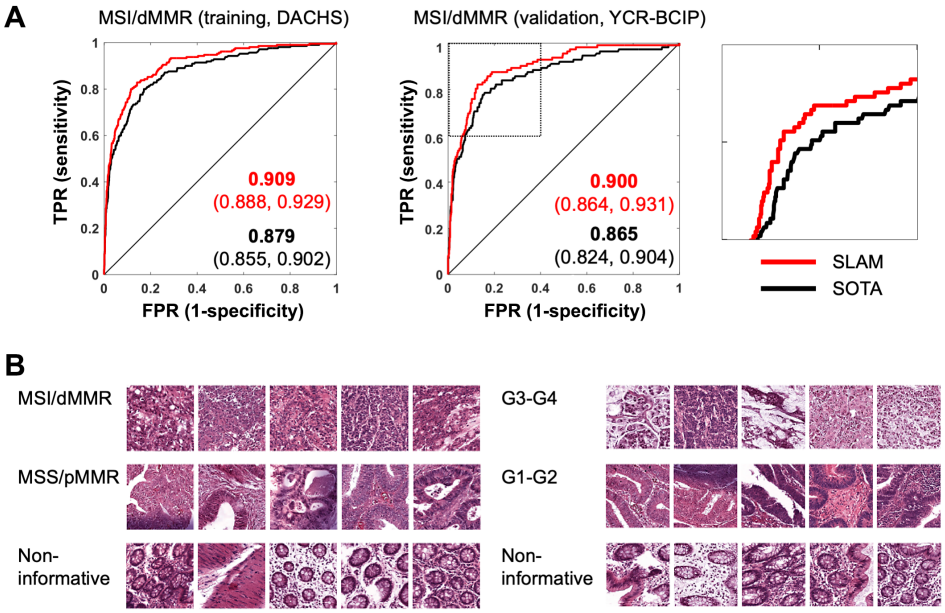

Figure 2. SLAM outperforms the SOTA model. (A) Patient-level accuracy shown as a ROC curve comparing SLAM (red) with SOTA (black). SLAM outperforms SOTA in the training cohort (DACHS, assessed via cross-validation) and in the external validation cohort (YCR-BCIP). The enlarged detail shows a performance gain in the high-sensitivity region of the ROC space. (B) Top predictive tiles from the top five patients in each group. For classification of MSI, MSS tumors, and normal tissue, as well as for high-grade (grade 3–4), low-grade (grade 1–2) tumors, and normal tissue, SLAM identifies histologically plausible image patches.

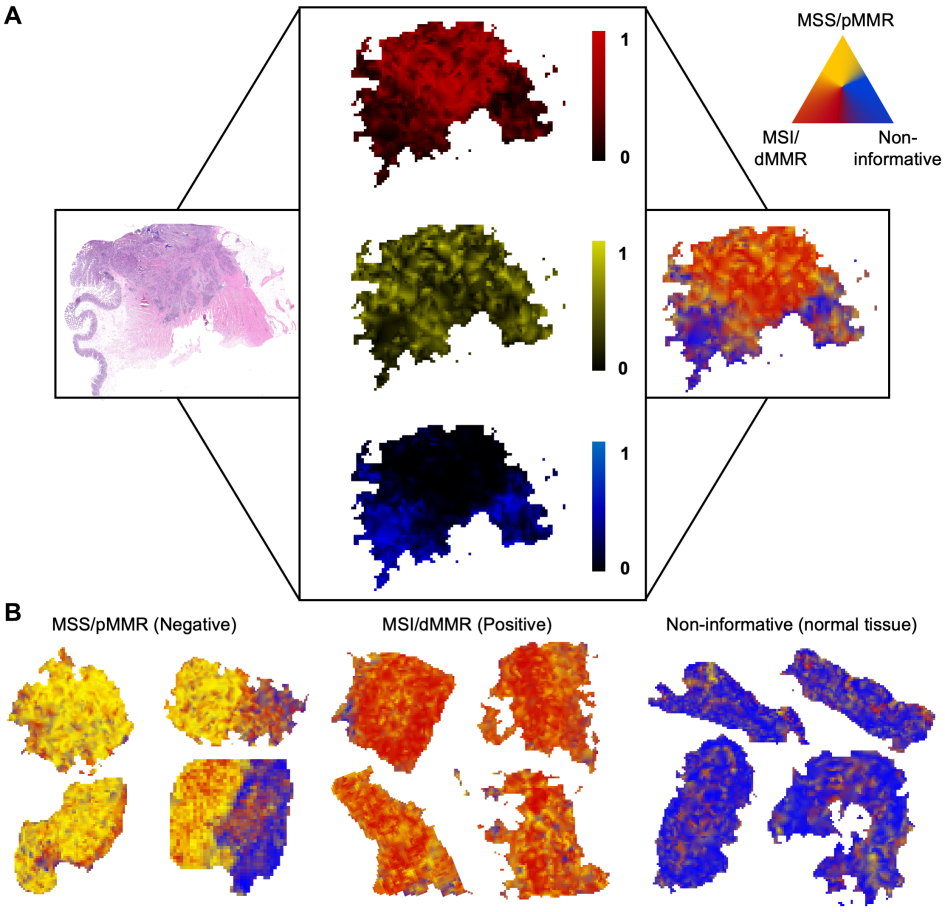

Figure 3. Visualizing multiplexed predictions. (A) Trivariate visualization of multiplexed spatially resolved predictions for tumor detection and subtyping. Based on tile-level predictions, three channels for each slide (MSI/dMMR, MSS/pMMR, and normal tissue) were created. These channels were merged to a single heatmap that contains each information. (B) Sample heatmaps of MSI/dMMR and MSS/pMMR tumors and normal tissue. In particular, the sample in the bottom right corner of MSS tumors shows that in some cases, a large area of non-tumor tissue is present on tumor slides.

© 2021 The Authors. *The Journal of Pathology* published by John Wiley & Sons, Ltd.
on behalf of The Pathological Society of Great Britain and Ireland. www.pathsoc.org

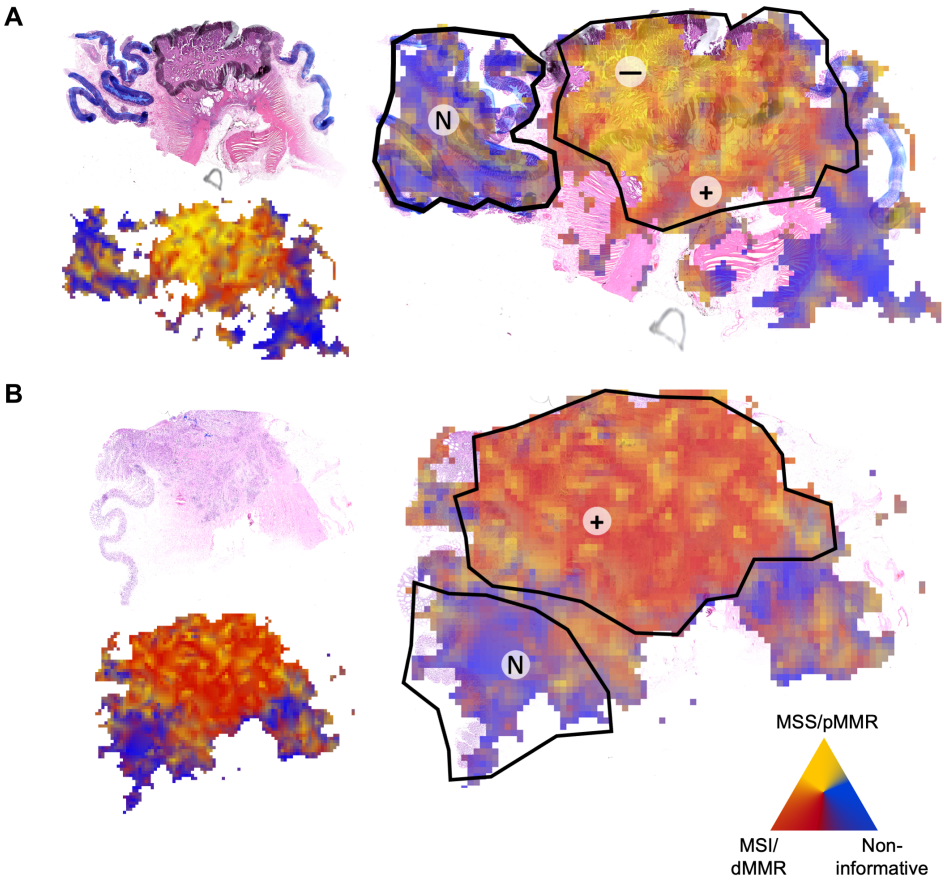

Figure 4. SLAM detects and excludes non-informative tissue and improves tumor subtyping. (A) Original WSIs and corresponding heatmap are shown for two representative cases. This is an MSS/pMMR tumor with surrounding healthy tissue. Although the patient was correctly assigned a high score for being MSS/pMMR, the trivariate visualization demonstrates only a slight heterogeneity of deep learning based within this patient. (−) denotes areas predicted to be negative (MSS/pMMR), (+) denotes areas predicted to be positive (MSI/dMMR), and (N) denotes normal tissue, as predicted by the model. (B) A representative patient for the MSI/dMMR class. In this case, the tumor tissue was homogeneously predicted to be MSI/dMMR.

baseline detection accuracy of the SOTA model with an AUROC of 0.590 (0.567, 0.623). SLAM achieved an AUROC of 0.609 (0.579, 0.623), showing just a slight improvement (see supplementary material, Figure S3B and Table S2). To assess the histopathologic plausibility of the proposed approach, we manually reviewed the 25 highest predictive tiles from the 25 highest predictive patients for all targets. We found that for prediction of MSI/dMMR status (Figure 2B), SLAM identified poorly differentiated and lymphocyte-rich image tiles as being the most predictive for MSI, whereas well-differentiated tumor glands with dirty necrosis were the most predictive for MSS. In the patches representative for normal tissue, i.e. non-informative tissue for the prediction of MSI/dMMR status, normal colon mucosa and smooth muscle tissue were the most prevalent tissue types. Similarly, for the model trained to predict tumor grading, high-grade and low-grade image tiles represented plausible tissue patterns (Figure 2B). Again, the highest scoring normal tissue tiles showed normal colon mucosa. Together, these results show SLAM's capabilities of improving prediction of genetic alterations in tumors by automatically detecting and excluding non-tumor tissue.

Multivariable visualization improves interpretability

Previous deep learning studies in digital pathology have provided univariate prediction heatmaps to make model predictions understandable to human observers. However, SLAM by design outputs multiplexed predictions, which require multivariate visualization. To achieve this, we developed a trivariate visualization method that allowed us to display tumor detection and predict genetic alterations in a single heat map (Figure 3A). Representative prediction maps for patients in each class are shown in Figure 3B. These heat maps provide assistance in a dissecting analysis of individual tumors as they display tumor heterogeneity in one glance. When reviewing the multiplexed visualization maps with expert observers, we found that tumor tissue in MSS/pMMR tumors and MSI/dMMR tumors (Figure 4A,B) could be localized by a human observer. In addition, trivariate visualization maps highlighted tumor heterogeneity. In true MSS/pMMR tumors, although the tumor tissue was overall visualized as 'yellow' (MSS/pMMR), the tumor invasive margin was occasionally (mis-)classified as MSI/dMMR (Figures 3B, 4A). Analysis of the underlying tissue slide revealed that these regions represented

© 2021 The Authors. *The Journal of Pathology* published by John Wiley & Sons, Ltd. on behalf of The Pathological Society of Great Britain and Ireland. www.pathsoc.org

the lymphocyte-rich tumor invasive margin, a well-known feature in CRC.

SLAM generalizes well to a large external cohort

As deep learning systems are prone to overfitting on the training dataset, external validation is an absolute requirement for assessing prediction performance. We externally validated the prediction performance for MSI/dMMR, training on all 2,448 DACHS cases and testing on 889 YCR-BCIP cases. For detection of MSI/dMMR status in YCR-BCIP, SLAM achieved an AUROC of 0.900 (0.864, 0.931) compared with a baseline achieved by the SOTA model of 0.865 (0.824, 0.904; Figure 1A). Correspondingly, accuracy, specificity, and F1 score were improved by SLAM compared with SOTA (supplementary material, Table S2). This demonstrates the generalizability of SLAM despite differences between the training set and the test set (e.g. a different method of determining the MSI/dMMR ground truth and the presence of pen marks in the training slides, but not in the slides in the validation set). In addition, a manual review of trivariate prediction maps showed that also in this cohort, tumor detection and subtyping was generally achieved in a spatially correct way.

Discussion

Tumor detection in digitized WSIs is a classical problem in computational pathology. A number of technical approaches to this problem were proposed even before the advent of deep learning methods [37]. Nowadays, deep learning approaches outperform hand-crafted pipelines for this problem [8]. However, in recent years a different type of problem has been increasingly addressed in computational pathology research. Beyond simple tasks, such as the detection of tumor tissue, it has been shown that deep learning is able to extract subtle visual features from histology images, making it possible to predict the presence of molecular alterations from routine pathology slides [21]. The central hypothesis to this approach is that the genotype gives rise to the phenotype, therefore genetic changes cause phenotypic changes and deep learning can infer the genotype of tumors just by observing tissue phenotype [22].

Here, we propose a simple workflow that improves prediction of genetic changes by simultaneously detecting tumor tissue in digitized pathology slides. Our approach only relies on slide-level labels, i.e. weak labels that are much cheaper and easier to generate than region-specific labels such as tile-level labels [24]. No manual tumor annotations whatsoever are required during training. We only trained on approximately 3,000 weakly labeled tissue slides, whereas previous studies have used much larger cohorts of up to 10,000 patients for training [23,24]. Although the tile-level labels in this approach are very noisy, we achieved a high performance for slide-level tumor detection (AUROC 0.980) and for molecular subtyping (AUROC 0.909 for MSI in the test cohort, 0.900 for MSI in the external validation cohort). In addition to accurate tumor detection on a slide level, our approach provides visualization maps for human readers that help in localizing tumor regions in heterogeneous tissue slides. These multiplexed visualization maps are generated with a new trivariate visualization method, which has previously only been applied for visualization of radiology image data [36]. This method allows expert observers to simultaneously check tumor localization capabilities and the predictions of molecular alterations of a deep learning model. We applied SLAM to multiple clinically relevant target features in CRC: MSI/dMMR status (which qualifies patients for immunotherapy [4]), *BRAF* mutational status (which qualifies patients for targeted therapy [6]), and grade of differentiation (an established histopathology feature defined on a case level). The cohort we used to investigate this was derived from a range of different pathology laboratories in southwest Germany, maximizing diversity of sample processing procedures. Finally, because all computational pathology methods should be validated in external cohorts in order to ensure generalizability [38], we evaluated classification performance on the YCR-BCIP cohort from 12 different institutions across the Yorkshire region of the UK. In this external cohort, we achieved a high performance with an AUROC 0.900 (0.864, 0.931) for MSI detection in addition to interpretable tumor localization. This demonstrates the robustness and generalizability of SLAM. Importantly, the idea behind SLAM is not to provide an automatic tool for perfect tumor segmentation in tissue slides, but to use detection of normal tissue as a tool to improve classification performance for the prediction of molecular alterations.

In an ecosystem of ever-increasing complexity of computational pathology workflows, the new approach provides a simple yet highly effective method for tumor localization and genotype prediction based on pathologic images. This simple method can be implemented using off-the-shelf models with transfer learning using standard deep learning libraries. Like any computational pathology method, before use in clinical routine, our method needs to undergo additional quality control and regulatory approval. A key limitation of our study was that it was only applied to a single tumor type, in which tumor tissue can be well distinguished from normal tissue. Future studies are needed to determine the performance of SLAM in other tumor types with more complex histopathologic patterns, such as pancreatic cancer or gastric cancer. We provide all of our source codes under an open-source license, allowing other groups to test SLAM in other disease contexts.

Acknowledgements

JNK is supported by the German Federal Ministry of Health (DEEP LIVER, ZMVI1-2520DAT111) and the Max-Eder-Programme of the German Cancer Aid (grant

© 2021 The Authors. *The Journal of Pathology* published by John Wiley & Sons, Ltd. on behalf of The Pathological Society of Great Britain and Ireland. www.pathsoc.org

#70113864). PQ is a National Institute for Health Research Senior Investigator. Histopathology and digital scanning were supported by a Yorkshire Cancer Research program grant L386. CT is supported by the German Research Foundation (DFG) (SFB CRC1382, SFB-TRR57). The DACHS study was supported by the German Research Council (BR 1704/6-1, BR 1704/6-3, BR 1704/6-4, CH 117/1-1, HO 5117/2-1, HO 5117/2-2, HE 5998/2-1, KL 2354/3-1, RO 2270/8-1 and BR 1704/17-1); the Interdisciplinary Research Program of the National Center for Tumor Diseases (NCT), Germany; and the German Federal Ministry of Education and Research (01KH0404, 01ER0814, 01ER0815, 01ER1505A, and 01ER1505B).

Author contributions statement

PLS, NGL and JNK conceptualized and designed the study. HB, JCC, EA, AB, MK, LRH, HIG, PQ, NPW and MH contributed tumor samples and corresponding molecular and clinical data. NGL, AE, EA, PLS and JNK preprocessed the data. MK and HIG provided histopathology expertise. DT, VS, TJB, DJ, CT and PQ provided additional computing and software resources. PLS, NGL and JNK performed the data analysis. PLS, HIG and JNK verified the underlying data. All authors had access to the underlying data. All authors contributed to the interpretation of the results. PLS and JNK wrote the first draft of the manuscript and all authors critically revised the manuscript. All authors approved the final version of the manuscript. All authors decided to submit this study and agreed to be accountable for all aspects of the work as recommended by the International Committee of Medical Journal Editors authorship criteria.

Data availability statement

All experiments were implemented in Matlab R2020b (Mathworks, Natick, MA, USA) and are freely available under an open-source license as the 'DeepHistology' package https://github.com/jnkather/DeepHistology. A re-implementation of the same algorithm in Python and PyTorch is available as the 'Histology Image Analysis' package at https://github.com/KatherLab/HIA. All results in this study were generated with the Matlab-based DeepHistology package.

References

1. Sung H, Ferlay J, Siegel RL, *et al*. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin* 2021; **71:** 209–249.
2. Benson AB, Venook AP, Al-Hawary MM, *et al*. Colon cancer, version 2.2021, NCCN clinical practice guidelines in oncology. *J Natl Compr Canc Netw* 2021; **19:** 329–359.
3. Molecular testing strategies for Lynch syndrome in people with colorectal cancer - NICE Guidance. [Accessed 1 September 2021]. Available from: https://www.nice.org.uk/guidance/dg27/chapter/1-Recommendations.
4. André T, Shiu KK, Kim TW, *et al*. Pembrolizumab in microsatellite-instability–high advanced colorectal cancer. *N Engl J Med* 2020; **383:** 2207–2218.
5. Greenson JK, Huang SC, Herron C, *et al*. Pathologic predictors of microsatellite instability in colorectal cancer. *Am J Surg Pathol* 2009; **33:** 126–133.
6. Kopetz S, Grothey A, Yaeger R, *et al*. Encorafenib, binimetinib, and cetuximab in *BRAF* V600E-mutated colorectal cancer. *N Engl J Med* 2019; **381:** 1632–1643.
7. Jiang Y, Yang M, Wang S, *et al*. Emerging role of deep learning-based artificial intelligence in tumor pathology. *Cancer Commun (Lond)* 2020; **40:** 154–166.
8. Echle A, Rindtorff NT, Brinker TJ, *et al*. Deep learning in cancer pathology: a new generation of clinical biomarkers. *Br J Cancer* 2021; **124:** 686–696.
9. Kather JN, Heij LR, Grabsch HI, *et al*. Pan-cancer image-based detection of clinically actionable genetic alterations. *Nat Cancer* 2020; **1:** 789–799.
10. Kather JN, Pearson AT, Halama N, *et al*. Deep learning can predict microsatellite instability directly from histology in gastrointestinal cancer. *Nat Med* 2019; **25:** 1054–1056.
11. Yamashita R, Long J, Longacre T, *et al*. Deep learning model for the prediction of microsatellite instability in colorectal cancer: a diagnostic study. *Lancet Oncol* 2021; **22:** 132–141.
12. Fu Y, Jung AW, Torne RV, *et al*. Pan-cancer computational histopathology reveals mutations, tumor composition and prognosis. *Nat Cancer* 2020; **1:** 800–810.
13. Binder A, Bockmayr M, Hägele M, *et al*. Morphological and molecular breast cancer profiling through explainable machine learning. *Nat Mach Intell* 2021; **3:** 355–366.
14. Krause J, Grabsch HI, Kloor M, *et al*. Deep learning detects genetic alterations in cancer histology generated by adversarial networks. *J Pathol* 2021; **254:** 70–79.
15. Woerl AC, Eckstein M, Geiger J, *et al*. Deep learning predicts molecular subtype of muscle-invasive bladder cancer from conventional histopathological slides. *Eur Urol* 2020; **78:** 256–264.
16. Diao JA, Wang JK, Chui WF, *et al*. Human-interpretable image features derived from densely mapped cancer pathology slides predict diverse molecular phenotypes. *Nat Commun* 2021; **12:** 1613.
17. Bychkov D, Linder N, Tiulpin A, *et al*. Deep learning identifies morphological features in breast cancer predictive of cancer ERBB2 status and trastuzumab treatment efficacy. *Sci Rep* 2021; **11:** 4037.
18. Foersch S, Eckstein M, Wagner DC, *et al*. Deep learning for diagnosis and survival prediction in soft tissue sarcoma. *Ann Oncol* 2021; **32:** 1178–1187.
19. Klein S, Quaas A, Quantius J, *et al*. Deep learning predicts HPV association in oropharyngeal squamous cell carcinomas and identifies patients with a favorable prognosis using regular H&E stains. *Clin Cancer Res* 2021; **27:** 1131–1138.
20. Chen M, Zhang B, Topatana W, *et al*. Classification and mutation prediction based on histopathology H&E images in liver cancer using deep learning. *NPJ Precis Oncol* 2020; **4:** 14.
21. Kather JN, Calderaro J. Development of AI-based pathology biomarkers in gastrointestinal and liver cancer. *Nat Rev Gastroenterol Hepatol* 2020; **17:** 591–592.
22. Coudray N, Ocampo PS, Sakellaropoulos T, *et al*. Classification and mutation prediction from non-small cell lung cancer histopathology images using deep learning. *Nat Med* 2018; **24:** 1559–1567.
23. Echle A, Grabsch HI, Quirke P, *et al*. Clinical-grade detection of microsatellite instability in colorectal tumors by deep learning. *Gastroenterology* 2020; **159:** 1406–1416.e11.
24. 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:** 1301–1309.

© 2021 The Authors. *The Journal of Pathology* published by John Wiley & Sons, Ltd. on behalf of The Pathological Society of Great Britain and Ireland. www.pathsoc.org

25. Perincheri S, Levi AW, Celli R, *et al*. An independent assessment of an artificial intelligence system for prostate cancer detection shows strong diagnostic accuracy. *Mod Pathol* 2021; **34:** 1588–1595.
26. Lu MY, Williamson DFK, Chen TY, *et al*. Data-efficient and weakly supervised computational pathology on whole-slide images. *Nat Biomed Eng* 2021; **5:** 555–570.
27. Lu MY, Chen TY, Williamson DFK, *et al*. AI-based pathology predicts origins for cancers of unknown primary. *Nature* 2021; **594:** 106–110.
28. Saillard C, Schmauch B, Laifa O, *et al*. Predicting survival after hepatocellular carcinoma resection using deep learning on histological slides. *Hepatology* 2020; **72:** 2000–2013.
29. Bilal M, Raza SEA, Azam A, *et al*. Novel deep learning algorithm predicts the status of molecular pathways and key mutations in colorectal cancer from routine histology images. *medRxiv* 2021. [Accessed 1 September 2021]. Available from: https://www.medrxiv.org/content/10.1101/2021.01.19.21250122v1 [Not peer reviewed].
30. Yamashita R, Long J, Banda S, *et al*. Learning domain-agnostic visual representation for computational pathology using medically-irrelevant style transfer augmentation. *arXiv [eessIV]* 2021. [Accessed 1 September 2021]. Available from: https://arxiv.org/abs/2102.01678 [Not peer reviewed].
31. Brenner H, Chang-Claude J, Seiler CM, *et al*. Long-term risk of colorectal cancer after negative colonoscopy. *J Clin Oncol* 2011; **29:** 3761–3767.
32. Amitay EL, Carr PR, Jansen L, *et al*. Association of aspirin and non-steroidal anti-inflammatory drugs with colorectal cancer risk by molecular subtypes. *J Natl Cancer Inst* 2019; **111:** 475–483.
33. Taylor J, Wright P, Rossington H, *et al*. Regional multidisciplinary team intervention programme to improve colorectal cancer outcomes: study protocol for the Yorkshire Cancer Research Bowel Cancer Improvement Programme (YCR BCIP). *BMJ Open* 2019; **9:** e030618.
34. Macenko M, Niethammer M, Marron JS, *et al*. A method for normalizing histology slides for quantitative analysis. In *2009 IEEE International Symposium on Biomedical Imaging: From Nano to Macro*. Boston, MA, USA: IEEE, 2009; 1107–1110.
35. Zhang X, Zhou X, Lin M, *et al*. ShuffleNet: an extremely efficient convolutional neural network for Mobile devices. *arXiv [csCV]* 2017. [Accessed 1 September 2021]. Available from: https://arxiv.org/abs/1707.01083 [Not peer reviewed].
36. Kather JN, Weidner A, Attenberger U, *et al*. Color-coded visualization of magnetic resonance imaging multiparametric maps. *Sci Rep* 2017; **7:** 41107.
37. Kather JN, Weis CA, Bianconi F, *et al*. Multi-class texture analysis in colorectal cancer histology. *Sci Rep* 2016; **6:** 27988.
38. Sobhani F, Robinson R, Hamidinekoo A, *et al*. Artificial intelligence and digital pathology: opportunities and implications for immuno-oncology. *Biochim Biophys Acta Rev Cancer* 2021; **1875:** 188520.

SUPPLEMENTARY MATERIAL ONLINE

Figure S1. Sample processing flowcharts for all targets

Figure S2. Detailed visual explanation of SLAM compared to SOTA

Figure S3. Receiver operating curves for additional targets

Figure S4. Confusion matrices for SLAM compared to SOTA

Table S1. Technical details of SLAM

Table S2. Overview of all results

© 2021 The Authors. *The Journal of Pathology* published by John Wiley & Sons, Ltd. on behalf of The Pathological Society of Great Britain and Ireland. www.pathsoc.org