Medical Image Analysis 79 (2022) 102474

Contents lists available at ScienceDirect

Medical Image Analysis

journal homepage: www.elsevier.com/locate/media

Benchmarking weakly-supervised deep learning pipelines for whole slide classification in computational pathology

Narmin Ghaffari Laleh[a], Hannah Sophie Muti[a], Chiara Maria Lavinia Loeffler[a], Amelie Echle[a], Oliver Lester Saldanha[a], Faisal Mahmood[b], Ming Y. Lu[b], Christian Trautwein[a], Rupert Langer[d], Bastian Dislich[c], Roman D. Buelow[e], Heike Irmgard Grabsch[f,g], Hermann Brenner[h,i,j], Jenny Chang-Claude[k,l], Elizabeth Alwers[h], Titus J. Brinker[m], Firas Khader[n], Daniel Truhn[n], Nadine T. Gaisa[e], Peter Boor[e], Michael Hoffmeister[h], Volkmar Schulz[o,p,q,r], Jakob Nikolas Kather[a,g,s,*]

[a] *Department of Medicine III, University Hospital RWTH Aachen, Aachen, Germany*
[b] *Department of Pathology, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA.*
[c] *Institute of Pathology, University of Bern, Switzerland.*
[d] *Institute of Pathology and Molecular Pathology, Kepler University Hospital, Johannes Kepler University Linz, Linz, Austria*
[e] *Institute of Pathology, University Hospital RWTH Aachen, Aachen, Germany*
[f] *Department of Pathology, GROW School for Oncology and Developmental Biology, Maastricht University Medical Center+, Maastricht, The Netherlands.*
[g] *Division of Pathology and Data Analytics, Leeds Institute of Medical Research at St James's, University of Leeds, Leeds, UK*
[h] *Division of Clinical Epidemiology and Aging Research, German Cancer Research Center (DKFZ), Heidelberg, Germany*
[i] *Division of Preventive Oncology, German Cancer Research Center (DKFZ) and National Center for Tumor Diseases (NCT), Heidelberg, Germany*
[j] *German Cancer Consortium (DKTK), German Cancer Research Center (DKFZ), Heidelberg, Germany*
[k] *Division of Cancer Epidemiology, German Cancer Research Center (DKFZ), Heidelberg, Germany*
[l] *Cancer Epidemiology Group, University Cancer Center Hamburg, University Medical Center Hamburg-Eppendorf, Hamburg, Germany*
[m] *Digital Biomarkers for Oncology Group, German Cancer Research Center (DKFZ), Heidelberg, Germany*
[n] *Department of Radiology, University Hospital RWTH Aachen, Aachen, Germany*
[o] *Department of Physics of Molecular Imaging Systems, Experimental Molecular Imaging, RWTH Aachen University, Aachen, Germany*
[p] *Fraunhofer Institute for Digital Medicine MEVIS, Bremen, Germany*
[q] *Comprehensive Diagnostic Center Aachen (CDCA), University Hospital Aachen, Aachen, Germany*
[r] *Hyperion Hybrid Imaging Systems GmbH, Aachen, Germany*
[s] *Else Kroener Fresenius Center for Digital Health, Medical Faculty Carl Gustav Carus, Technical University Dresden, Dresden, Germany*

ARTICLE INFO

Article history:
Received 20 August 2021
Revised 7 April 2022
Accepted 3 May 2022
Available online 4 May 2022

Keywords:
Computational pathology
Artificial intelligence
Weakly-supervised deep learning
Vision transformers
Convolutional neural networks
Multiple-Instance Learning

ABSTRACT

Artificial intelligence (AI) can extract visual information from histopathological slides and yield biological insight and clinical biomarkers. Whole slide images are cut into thousands of tiles and classification problems are often weakly-supervised: the ground truth is only known for the slide, not for every single tile. In classical weakly-supervised analysis pipelines, all tiles inherit the slide label while in multiple-instance learning (MIL), only bags of tiles inherit the label. However, it is still unclear how these widely used but markedly different approaches perform relative to each other.

We implemented and systematically compared six methods in six clinically relevant end-to-end prediction tasks using data from N=2980 patients for training with rigorous external validation. We tested three classical weakly-supervised approaches with convolutional neural networks and vision transformers (ViT) and three MIL-based approaches with and without an additional attention module. Our results empirically demonstrate that histological tumor subtyping of renal cell carcinoma is an easy task in which all approaches achieve an area under the receiver operating curve (AUROC) of above 0.9. In contrast, we report significant performance differences for clinically relevant tasks of mutation prediction in colorec-

Abbreviations: MIL, Multiple-Instance Learning; AI, Artificial Intelligence; ViT, Vision Transformer; WSI, Whole Slide Image; AUROC, Area Under the Receiver Operating Characteristic Curve; H&E, Hematoxylin and Eosin; CNN, Convolutional Neural Network; EBV, Epstein-Barr virus; MSI, Microsatellite Instability.

* Corresponding author:
E-mail address: jkather@ukaachen.de (J.N. Kather).

https://doi.org/10.1016/j.media.2022.102474
1361-8415/

tal, gastric, and bladder cancer. In these mutation prediction tasks, classical weakly-supervised workflows outperformed MIL-based weakly-supervised methods for mutation prediction, which is surprising given their simplicity. This shows that new end-to-end image analysis pipelines in computational pathology should be compared to classical weakly-supervised methods. Also, these findings motivate the development of new methods which combine the elegant assumptions of MIL with the empirically observed higher performance of classical weakly-supervised approaches. We make all source codes publicly available at https://github.com/KatherLab/HIA, allowing easy application of all methods to any similar task.

Introduction

The state of the art in end-to-end computational pathology

Artificial intelligence (AI)-based image analysis is widely used for end-to-end classification of histopathological whole slide images (WSI). Common applications of such end-to-end workflows are tumor detection (Campanella et al., 2019, Pinckaers et al., 2021), subtyping (Lu et al., 2021, Wang et al., 2020, Zhu et al., 2021), and grading (Bulten et al., 2020, Shaban et al., 2020, Ström et al., 2020). In these tasks, an image analysis pipeline recapitulates, automates, and potentially improves pathologists' assessment of WSI. However, AI has also been used to perform image analysis tasks that exceed human capabilities, including prediction of molecular alterations (Kather et al., 2020), survival (Skrede et al., 2020, Yamamoto et al., 2019), and treatment response (Echle et al., 2020) directly from routine WSI. Collectively, these broad applications of AI in WSI image analysis are termed "computational pathology" and widespread clinical adoption is ultimately expected once routine diagnostic workflows are fully digitalize (Kather and Calderaro, 2020, Rony et al., 2019) (Figure 1**A)**. Glass slides stained with hematoxylin and eosin (H&E) are ubiquitously available for almost every cancer patient. Hence, AI methods are expected to easily integrate with existing diagnostic pathways, improving outcomes and providing cost savings (Kacew et al., 2021).

Current limitations: a lack of unbiased benchmarking studies

However, a major limitation for the development, validation, and commercialization of computational pathology methods is the lack of systematic comparison (i.e. benchmarking) of different technologies. While the earliest studies in 2018 employed a weakly-supervised approach based on a convolutional neural network (CNN) and spatial averaging (Coudray et al., 2018), recent studies have proposed conceptually new technologies, including multiple-instance learning (Campanella et al., 2019, Lu et al., 2021, Lu et al., 2021) with or without attention-based aggregation functions (Saillard et al., 2020, Schirris et al., 2021). In addition, computational pathology is an applied field that follows trends in basic computer vision research. Thus, it can be anticipated that classical CNN architectures such as ResNets (Residual Neural network) will be ultimately replaced by more powerful and efficient CNNs such as EfficientNet (Tan and Le, 2019) or non-convolutional AI approaches such as Vision Transformers (ViT) (Dosovitskiy et al., 2020). However, for academic and commercial actors in the field of computational pathology, choosing the best method for an end-to-end problem is currently not easily possible. On a conceptual level, there is no systematic evidence on which methods yield the best performance for clinically relevant problems. This prevents researchers, pathologists, and companies from making optimal design choices for a computational pathology application. On a practical level, there is currently no implementation of the whole spectrum of AI methods for computational pathology.

Aim of the present study

In the present study, we systematically collected WSI datasets for six clinically common end-to-end prediction tasks with diagnostic or therapeutic relevance. In renal cell carcinoma, we investigated the classification of morphological subtypes, which is a widely studied problem (Lu et al., 2021). In colorectal cancer, we investigated AI-based prediction of the immunotherapy biomarker microsatellite instability (MSI) (Bilal et al., 2021, Echle et al., 2020, Echle et al., 2021, Kather et al., 2019) and mutations in the BRAF gene, which is a directly targetable genetic alteration (Bilal et al., 2021, Kather et al., 2020, Kopetz et al., 2019, Schrammen et al., 2021). In gastric cancer, we investigated the prediction of established or potential biomarkers for immunotherapy MSI and Epstein-Barr virus (EBV) positivity (Muti et al., 2021). Finally, in bladder cancer, we investigated the prediction of FGFR3 mutational status, which is a clinically approved therapeutic target (Loeffler et al., 2021). For each of these tasks, we presented datasets from two different institutions, allowing us to provide a benchmark with external validation (Figure 1 B-E, Suppl Figures 1-4). We aimed to be unbiased in terms of methods selection for this benchmark study. Therefore, we identified the most commonly used image analysis approaches via a systematic review of the literature. We included all approaches for weakly-supervised end-to-end slide-level classification tasks (Suppl. Figure 5) and subsequently benchmarked them on the multi-tumor datasets.

Methods

Ethics statement and patient cohorts

All experiments were conducted in accordance with the Declaration of Helsinki. For this study, we used anonymized H&E stained slides obtained from formalin-fixed paraffin-embedded (FFPE) material from the "The Cancer Genome Atlas" (TCGA) archive (available at https://portal.gdc.cancer.gov), a large, multi-centric collection of tissue specimens obtained from multiple hospitals across different countries (Cancer Genome Atlas Research Network, 2014a, Cancer Genome Atlas Research Network, 2014b; Ricketts et al., 2018). In addition, we used four proprietary datasets: the DACHS study ("Darmkrebs: Chancen der Verhütung durch Screening"), 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 coordinated by the German Cancer Research Center (Heidelberg, Germany) (Brenner et al., 2011, Hoffmeister et al., 2020, Brenner et al., 2006) and supported by the the NCT Tissue Bank at the Institute of Pathology, University of Heidelberg. The DACHS study was approved by the ethics committees of the University of Heidelberg and of the Medical Chambers of Baden-Württemberg and Rhineland-Palatinate, and all participants signed an informed consent. The BERN dataset is a single-center dataset collected from clinical routine samples at the pathology archive at Inselspital, University of Bern (Bern,

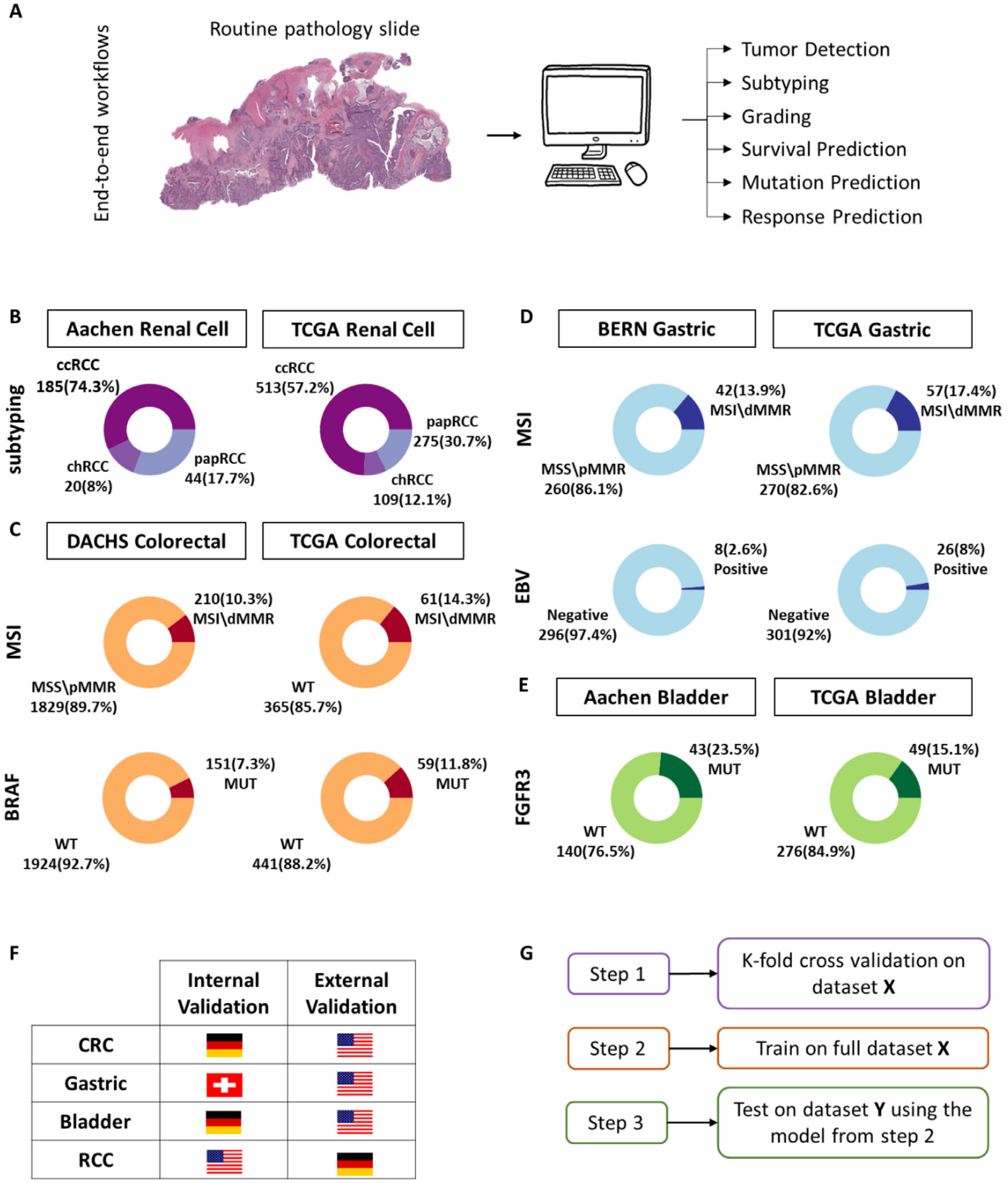

Fig. 1. Outline of this study. A) End-to-end artificial intelligence (AI) methods in computational pathology are used to predict a range of features. B) Patient cohorts for renal cell carcinoma, C) for colorectal cancer, D) for gastric cancer and E) for bladder cancer. F) Country of origin of all cohorts. G) Experimental design in this study.

Switzerland) (Dislich et al., 2020). The use of this data set was approved by the local ethics commission, specifically granting the use of archival tissue for molecular and immunohistochemical analysis as well as tissue microarray construction (University of Bern, Switzerland, no. 200/14). The use of archival tissue from this cohort for molecular analysis was approved by the local ethical commission (Technical University of Munich, No. 2136/08). Similarly, the AACHEN-RCC dataset and the AACHEN-BLADDER datasets originated from a single high-volume medical center, the pathology archive at RWTH Aachen University Hospital (Aachen, Germany). The collection of patient samples from Aachen was approved by the local Ethics board (AACHEN-RCC: EK315/19, AACHEN-BLADDER: EK455/20). All cohorts were anonymized at the time of analysis. Suppl. Table 1 shows patient numbers and a clinico-pathological description of all cohorts.

Identification of commonly used weakly-supervised prediction pipelines

We used the search term “((deep learning) AND ((digital histopathology) OR (whole slide)) AND (cancer))” for a structured search of the PubMed database (https://pubmed.gov). To follow the main goal of the current study, we selected the studies which aimed to predict clinically relevant targets from WSIs using

Table 1
Hyperparameters and technical details for all approaches.

	Hyperparameters and architecture	Reference, technical	Reference, medical
ResNet	•Resnet-18 pre-trained on ImageNet (18 layers, the last layer was changed from 1000 output neurons to N output neurons for N classes) •batch size = 32 •Maximum number of tiles = 500 •Optimizer = Adam •learning rate = 1e-5 (weight decay = 1e-5) •Freeze ratio of layers =0.5 (weights and biases in the first 9 layers were not trainable, weights and biases in the last 9 layers were trainable)	(He et al., 2016)	(Fu et al., 2020, Kather et al., 2020, van Treeck et al., 2021)
EfficientNet	•Pre-trained on ImageNet (efficientnet-b7) •batch size = 32 •Maximum number of tiles = 500 •Optimizer = Adam •learning rate = 1e-4 (weight decay = 1e-5) •Freeze ratio of layers =0.25	(Tan and Le, 2021)	(Bengs et al., 2021)
ViT	•Pre-trained on ImageNet (B_32_imagenet1k, 24 layers, the last layer was changed from 1000 output neurons to N output neurons for N classes) •batch size = 32 •Maximum number of tiles = 500 •Optimizer = Adam •learning rate = 1e-4 (weight decay = 1e-5)	(Dosovitskiy et al., 2020, Touvron et al., 2020)	N/A
MIL	•Extract features using a Resnet-50 (which is pre-trained on ImageNet) for all the tiles of a slide •batch size = 1 •Optimizer = Adam •learning rate = 1e-4 (weight decay = 1e-5) •dropout = True	(Dietterich et al., 1997)	(Campanella et al., 2019)
AttMIL	•Extract features using a Resnet-50 (which is pre-trained on ImageNet) for all the tiles of a slide •batch size = 32 •Optimizer = Adam •learning rate = 1e-3 (weight decay = 1e-5)	(Ilse et al., 2018)	(Yao et al., 2020)
CLAM	•Extract features using a Resnet-50 (which is pre-trained on ImageNet) for all the tiles of a slide •batch size = 1 •Optimizer = Adam •learning rate = 1e-5 (weight decay = 1e-5) •Model size = small •Bag loss = Cross Entropy •Instance Loss = SmoothTop1SVM Berrada et al., 2018. •dropout = True	(Lu et al., 2021)	(Lu et al., 2021)

weakly-supervised end-to-end classification approaches. We excluded all studies which were designed for segmentation and object detection in WSIs, for none-histopathology images, none-H&E stained images, survival, and prognostication. We sub-classified the studies into classical weakly-supervised, multiple-instance learning (MIL) and other methods (Suppl. Figure 5).

Prediction tasks and experimental design

We benchmarked all technical approaches in six end-to-end prediction tasks, selected to represent a wide range of clinically relevant problems. To this end, we trained algorithms to predict each of these targets from raw histological whole slide images: (1) Diagnosis of renal cell carcinoma subtype (clear cell RCC, chromophobe RCC, and Papillary RCC); (2) prediction of microsatellite instability (MSI) or mismatch repair deficiency (dMMR) in colorectal cancer; (3) prediction of BRAF mutation in colorectal cancer; (4) prediction of MSI/dMMR in gastric cancer; (5) detection of Epstein-Barr Virus (EBV) presence in gastric cancer and (6) prediction of FGFR3 point mutations in bladder cancer. MSI and dMMR have a very high degree of overlap and are interchangeably used in clinical routines (Molecular testing strategies for Lynch syndrome in people with colorectal cancer - NICE Guidance., 2019). Here, we use the term “MSI” throughout the study. For each task, one training and one testing cohort were defined. First, we performed a within-cohort experiment on each training set by patient-level three-fold cross-validation (DACHS-CRC, BERN-Gastric, AACHEN-Bladder, TCGA-RCC). Subsequently, we re-trained a classifier for each prediction task on the training cohorts and externally validated it on the test cohort (TCGA-CRC, TCGA-Gastric, TCGA-Bladder, AACHEN-RCC). The validation cohorts were not used for any other purpose except for the validation of the final model.

Ground truth for prediction tasks

The ground truth for the prediction targets were obtained as follows: For TCGA-CRC and TCGA-STAD, MSI and EBV status were obtained from a public source (Liu et al., 2018) as described before (Kather et al., 2020). For TCGA-RCC, images from the three morphological subtypes were obtained separately from the GDC data portal (TCGA-KIRP for papillary, KIRC for clear cell, and KICH for chromophobe tumors). In DACHS, MSI status was obtained by 3-plex PCR, and BRAF V600E mutational status was obtained by immunohistochemistry (IHC) on tissue microarrays and by Sanger sequencing, as described before (Alwers et al., 2019, Jia et al., 2016). For the BERN cohort, MSI/dMMR status was obtained with IHC for DNA repair enzymes and EBV status was obtained by Epstein-Barr virus (EBV)-encoded RNA (EBER) in-situ hybridization. AACHEN-

BLADDER comprised bladder carcinomas from a real-world cohort (Loeffler et al., 2021) and FGFR3 mutational status was obtained by whole-exome sequencing or identified using the SNaPshot method (Hurst et al., 2009). In the AACHEN-RCC cohort, the subtype was retrieved from the routine pathology report.

Image preprocessing

The input images for all the methods were preprocessed based on the “Aachen protocol for Deep Learning histopathology” (Muti et al., 2020). Based on this protocol, the digitized whole slide images were tessellated into smaller image tiles of (512 ×512) pixels at a resolution of 0.5 micrometers per pixel (MPP). During this process, tiles containing background and artifacts were removed from the data set (using canny edge detection in Python’s OpenCV package). Extracted tiles were color normalized using the Macenko method to reduce the inter-cohort color bias (Macenko et al., 2009). No manual annotations were applied to the whole slide images and all models were trained only with slide-level labels.

Artificial intelligence methods

For our benchmarking task, we implemented and systematically compared six different methods for end-to-end artificial intelligence on WSI (Table 1). “Classical weakly-supervised” methods assume that all tiles from a given slide inherit the slide label for classification (Coudray et al., 2018; Kather et al., 2019).Models are trained on N randomly selected tiles per WSI and tile-level predictions are averaged for each patient. Empirically, this can yield clinical-grade performance despite weak labels (Coudray and Tsirigos, 2020, Echle et al., 2020), even without any annotation (Kather et al., 2020, Muti et al., 2021). Three different AI models were used within this classical approach: ResNet, EfficientNet, and Vision Transformers (ViT).

1. ResNets are currently the de-facto standard for supervised transfer learning due to their higher performance and efficiency when compared to other CNN models (He et al., 2016). The model was pre-trained on ImageNet and fine-tuned by transfer learning on each benchmark task separately. This approach was motivated by a number of previous studies (Echle et al., 2020, Kather et al., 2019, Muti et al., 2021, van Treeck et al., 2021).
2. EfficientNet aims to scale up the baseline CNN which has been referred to as EfficientNet-B0 (Tan and Le, 2019). The common approach in designing any CNN is to develop a smaller version of the network and then scale it up to reach higher performance. EfficientNet scales the width, depth, and resolution of the network using the compound scaling method, which achieves state-of-the-art accuracies on smaller and therefore faster networks.
3. ViT is the most modern AI architecture analyzed within the classical workflows. Since 2017, attention-based models have become the dominant selection in natural language processing (NLP) (Vaswani et al., 2017). In 2020, the high performance of transformers in visual tasks has been demonstrated (Dosovitskiy et al., 2020). The input to the vision transformer is flattened 2D patches extracted from the original image. All the layers of the transformer use a constant latent vector size. Through a patch embedding block, the flattened patches get mapped to D dimensions using a trainable linear projection. This step is followed by a position embedding block which adds positional information to each patch. The encoder of the transformer consists of alternating layers of self-attention, multilayer perceptron (MLP), layer norm (LN) before each block, and residual connections after each block. Although ViTs showed very good performance on the ImageNet data set, their performance on histopathological images with smaller sizes has not been systematically investigated before this study.

The conceptual limitations of the classical weakly-supervised computational pathology workflow are addressed by multiple instance learning (MIL). MIL groups all tiles from a given patient in “bags”. The label of individual tiles is unknown, but the label of the bag is positive if there is at least one positive instance within that bag. In theory, MIL is well suited to handle a heterogeneous set of tiles obtained from different regions in a WSI. In this study, we tested three different MIL methods: Classical MIL, Attention-based MIL (AttMIL), and Clustering constrained Attention MIL (CLAM).

1. Classical MIL has been used diversely in the processing of histopathological images to address the problem of label inheritance from slides to tiles (Das et al., 2018, Ilse et al., 2018, Sudharshan et al., 2019, Xu et al., 2014) The basic framework of MIL was in the past successfully applied to large-scale image classification tasks in histopathology (Campanella et al., 2019). The naive approach uses a max-pooling layer, so that the patch with the highest predicted probability score for the positive class is used to represent the final slide-level prediction.
2. AttMIL uses an attention mechanism consisting of two fully connected layers which compute a scalar attention score for each tile. Each of the tiles’ embeddings is then scaled with the softmax of the tile’s attention score. By summing up these scaled embeddings, we obtain a bag-level feature vector. Another fully connected layer then transforms this bag-level feature vector into a final classification. A subset of each patient’s tiles is considered sufficient. In each epoch, the tiles are resampled, enabling the model to be trained on multiple patients in each batch (Ilse et al., 2018).
3. CLAM has been designed initially to overcome the challenges in the standard MIL approaches (Lu et al., 2021). By using an attention-based pooling layer, it is able to detect the most informative regions on a WSI. CLAM was empirically shown to outperform classical MIL (Lu et al., 2021). Compared to standard MIL methods, which use the gradient signal only from one single instance from each bag to update the learning parameters, CLAM aggregates patch-level features into slide-level information required for classification, in theory achieving higher robustness. CLAM uses low-dimensional features extracted from the input tiles (which is computationally expensive), but the actual training only uses feature vectors and the required computational power and time for training of this model is very low. The source code for CLAM and MIL methods are taken from https://github.com/mahmoodlab/CLAM and were modified based on our workflow. Figure 2 shows the workflow for each model.

For training of all the methods, we used early stopping based on AUROC with at least 5 (classical methods) or 20 (MIL-based methods) with the patience of 5 epochs. The reason for this selection is that in all classical weakly-supervised methods, models are pre-trained ImageNet. However, in MIL-based weakly-supervised methods, features are extracted with a ResNet model trained on ImageNet, but the classification model is initialized with random weights.

Hyperparameter tuning

For hyperparameter tuning, we used MSI prediction in the DACHS-CRC cohort (70% train, 30% test) and fine-tuned the required hyperparameters for each specific method. For classical weakly-supervised methods and MIL-based methods, we used a minimum training epoch of 10 (50, respectively) with a patience

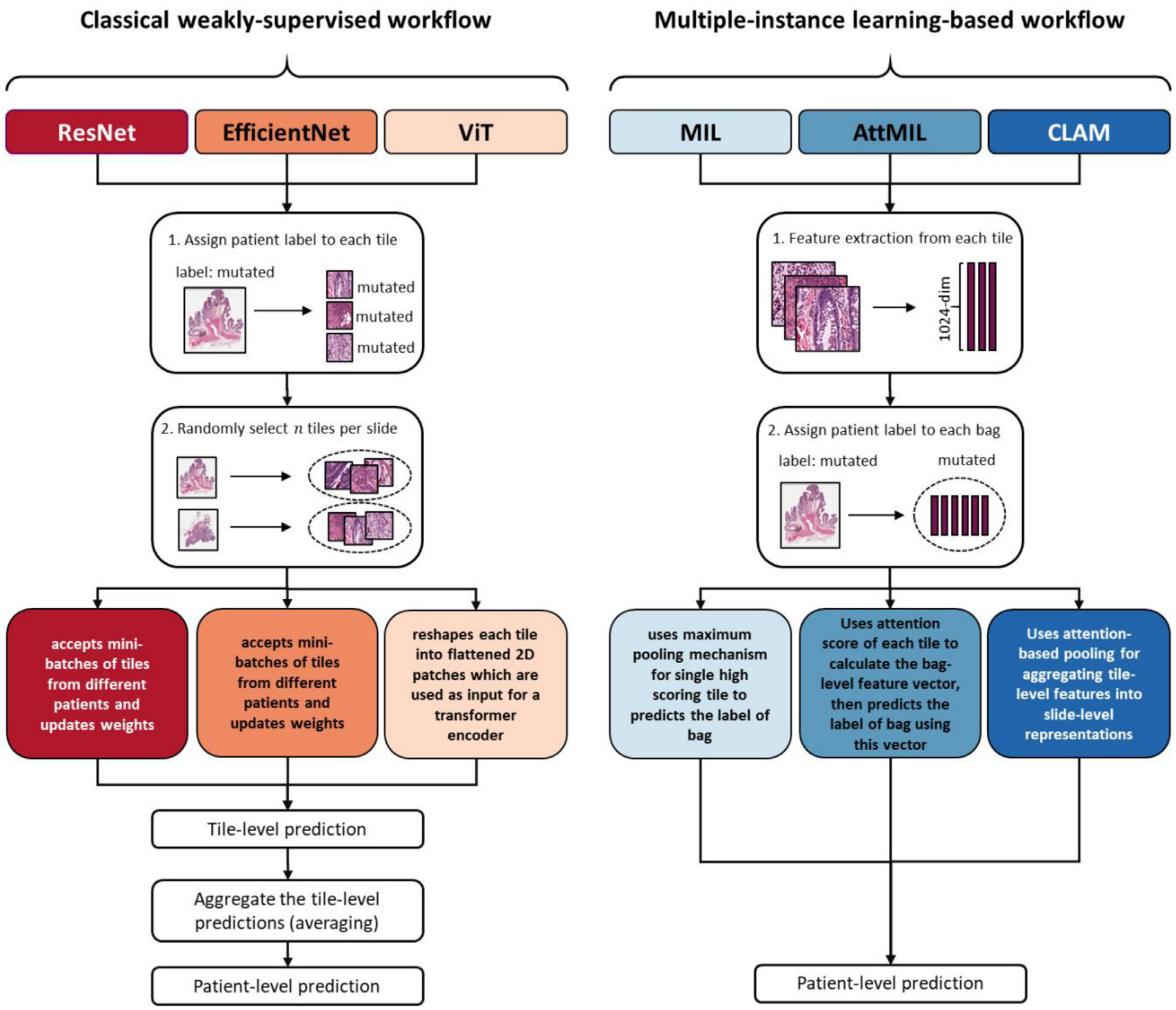
Classical weakly-supervised workflow
Multiple-instance learning-based workflow
ResNet
EfficientNet
ViT
MIL
AttMIL
CLAM
1. Assign patient label to each tile
label: mutated
mutated
mutated
mutated
1. Feature extraction from each tile
1024-dim
2. Randomly select n tiles per slide
2. Assign patient label to each bag
label: mutated
mutated
accepts mini-batches of tiles from different patients and updates weights
accepts mini-batches of tiles from different patients and updates weights
reshapes each tile into flattened 2D patches which are used as input for a transformer encoder
uses maximum pooling mechanism for single high scoring tile to predicts the label of bag
Uses attention score of each tile to calculate the bag-level feature vector, then predicts the label of bag using this vector
Uses attention-based pooling for aggregating tile-level features into slide-level representations
Tile-level prediction
Aggregate the tile-level predictions (averaging)
Patient-level prediction
Patient-level prediction

Fig. 2. Schematic workflow of the methods. ResNet and EfficientNet as well as Vision Transformers (ViT) were used for weakly-supervised end-to-end prediction benchmark tasks. In addition, classical multiple instance learning (MIL) and attention-based MIL (AttMIL), and clustering-constrained attention multiple-instance learning (CLAM) were used for the same tasks. While classical workflows use different models (ResNet, EfficientNet, ViT), they all cast slide labels to image tiles. In contrast, MIL, AttMIL, CLAM cast slide labels to bags of image tiles without assuming that every single tile reflects the target of interest.

of 5 and then stopped training if the validation loss did not decrease. For all methods, we trained models for three different learning rates (10e-3, 10e-4, and 10e-5), and for ResNet and EfficientNet, we searched for the best freeze ratio (0.25, 0.5, and 0.75). These resulted in the optimum learning rate for each method, ResNet (LR=10-5, freeze ratio=0.5), EfficientNet (LR=10e-4, freeze ratio = 0.25), ViT (LR =10e-4), MIL (LR = 10e-4), AttMIL (LR=10-3) and CLAM (LR=10e-5), which were used for all subsequent experiments.

Statistics

The primary statistical endpoint was the area under the receiver operating curve (AUROC) calculated on the level of patients. Confidence intervals were obtained by 1000x bootstrapping based on sampling with replacement. For binary classification tasks, AUROCs were identical for both groups, and therefore, only the AUROC for the positive group is reported. For multiclass classification tasks, we binarized the ground truth labels (for each class) and calculated the AUROC for the prediction scores of the same class (macro-averaging). To quantify whether performance differences between models were statistically significant, we used DeLong's method. This method tests whether two models have a significant difference in their performance and accounts for the role of randomness in finite datasets (DeLong et al., 1988). The output of this method is the z score (difference of AUROC of the output performance of two models divided by its standard error) and the p-value.

Code availability

All methods are implemented using Python 3.8 with PyTorch and all source codes for preprocessing are available at https://github.com/KatherLab/preProcessing and all codes for training and evaluating the models with the Histology Image Analysis package are available at https://github.com/KatherLab/HIA under an open-source license.

Results

Identification of commonly used weakly-supervised prediction pipelines

For this benchmark study, we strived to use a representative selection of widely used weakly-supervised end-to-end prediction pipelines for computational pathology. We deliberately focused on studies indexed in PubMed to include methods which have been applied to medical research questions. This literature search yielded 548 results (Suppl. Figure 5) of which we excluded 452 (82%) as they were not aimed at predicting clinically relevant targets from digital WSIs using weakly-supervised end-to-end approaches. This resulted in 96 studies, of which 65 used classical weakly-supervised workflows (Jang et al., 2021, Kanavati et al., 2020, Schrammen et al., 2021, Wang et al., 2021), 10 used multiple-instance learning-based workflow (Campanella et al., 2019, Chen et al., 2021, Lu et al., 2021, Sharma et al., 2021) and 21 used other methods (Li et al., 2021, Wang et al., 2020). Furthermore, we classified the other methods, into multi-field (resolution) convolutional neural networks (3 studies) (Kosaraju et al., 2020, Sha et al., 2019), multi-step CNN approaches (11 studies) (Yamashita et al., 2021), and the combination of CNN with machine learning schemes like support vector machine (SVM) (7 studies) (Gheisari et al., 2018, Kott et al., 2021). Based on these findings, we selected three classical weakly-supervised approaches (ResNet, EfficientNet, and ViT-based) and three multiple-instance learning (MIL)-based approaches (CLAM, classical MIL, and AttMIL) for this study. For all approaches, we optimized the learning rate and freeze ratio (Suppl. Figure 6).

All methods achieve high performance for subtyping of renal cell carcinoma

Morphological subtyping of renal cell carcinoma (RCC) into the clear cell, chromophobe, and papillary subtypes is widely studied and clinically relevant. Using the “The Cancer Genome Atlas” (TCGA) cohort (TCGA-RCC, N=897 patients, Suppl. Table 1), we benchmarked classification performance of end-to-end prediction workflows based on ResNet, EfficientNet, and ViT as well as classical MIL, AttMIL, and CLAM (Figure 2). We found that in stratified three-fold cross-validation, all methods achieved a high classification performance with an area under the receiver operating curve (AUROC) values above 0.90 (Table 2). EfficientNet achieved the highest absolute performance with AUROCs of 0.983 (with 95% confidence interval of 0.975 - 0.990), 0.992 (0.987 - 0.997) and 0.986 (0.980 - 0.992) for detection of all three classes. The weakly-supervised ViT-based approach yielded AUROCs of 0.977 (0.967 - 0.985), 0.984 (0.970 - 0.994) and 0.985 (0.979 - 0.991), demonstrating the efficiency of simple classical methods. While MIL-based methods yielded a high absolute performance, this was consistently lowest in all target classes, with MIL achieving AUROCs of 0.951 (0.935 - 0.964), 0.955 (0.934 - 0.971) and 0.943 (0.925 - 0.959). Next, we trained classifiers on all TCGA cases and validated them on our in-house dataset (N=249 patients). As expected, performance values slightly decreased, but classic weakly-supervised methods remained the highest-scoring approaches with for example AUROCs of 0.971 (0.952 - 0.986), 0.949 (0.897 - 0.989), and 0.980 (0.963 - 0.994) for all classes in the ResNet-based approach (Table 3, Suppl. Figure 7). Similarly, areas under the precision-recall curve (AUPRCs, Table 4) values also show a high performance for all three classes. As an example, CLAM reaches AUPRCs of 0.989 (0.981-0.996), 0.829 (0.652-0.954), and 0.843 (0.738-0.931), which based on the baseline values of (the ratio of positive cases to total number of samples) 0.74, 0.08, and 0.17 are very high. However, the performance differences between all methods compared to Resnet (Suppl.Table 2), EfficientNet (Suppl.Table 3), ViT (Suppl. Table 4), MIL (Suppl. Table 5), AttMIL (Suppl. Table 6), and CLAM (Suppl. Table 7) did not reach statistical significance in the external validation experiments. We conclude that AI-based RCC subtyping is achievable with almost perfect accuracy compared to the ground truth by any of the tested computational pathology methods.

Classic weakly-supervised methods excel in mutation prediction in colorectal cancer

Next, we focused on the prediction of clinically actionable genetic alterations directly from H&E histology WSI: MSI and BRAF in colorectal cancer, MSI and EBV in gastric cancer, and FGFR3 mutations in bladder cancer. In a cross-validated experiment in the large DACHS cohort of colorectal cancer, EfficientNet achieved a state-of-the-art AUROC of 0.930 (0.906 - 0.950; N=2039 patients). The classical ResNet-based approach achieved the second-highest performance with an AUROC of 0.917 (0.895 - 0.938). Classic weakly-supervised classifiers generalized well to the external validation cohort (TCGA-CRC, N=426 patients) with ViT and EfficientNet yielding the highest and second-highest performance for MSI prediction with AUROCs of 0.885 (0.834 - 0.926) and 0.883 (0.829 - 0.928), respectively. Compared to the other approaches MIL, AttMIL, and CLAM, the performance of ViT-based classical weak supervision was significantly higher (Suppl. Table 4). Although ViT slightly outperformed EfficientNet (z=0.06), their direct comparison did not reach statistical significance (p=0.95, Suppl. Table 4). All other methods, in particular, MIL-based methods reached much lower performances in within-cohort experiments, with classical MIL, AttMIL, and CLAM yielding AUROC of 0.709 (0.675 - 0.742), 0.880 (0.751 - 0.909), and 0.795 (0.763 - 0.828), respectively (Table 2). Likewise, in external validation experiments, MIL, AttMIL, and CLAM yielded the lowest performance (Table 3) which was statistically significantly inferior to all other approaches (p<=0.01, Suppl. Table 5,6, and 7). Prediction of BRAF mutational status (N=2075 patients in cross-validation) resulted in the same ranking of algorithms with EfficientNet achieving the highest (AUROC 0.856 [0.825 - 0.887]), and classical MIL achieving the lowest performance (AUROC 0.629 [0.580 - 0.676]). Also in external validation (Table 3), EfficientNet (AUROC 0.808 [0.749 - 0.861]) and ViT (0.786 [0.719 - 0.844]) significantly (p<=0.02) outperformed MIL and AttMIL approaches (Suppl. Table 3 and Suppl. Table 4). To check the effect of imbalanced sample size for performance evaluation, AUPRC values were calculated for external validation experiments (Table 4). In line with our previous findings, ViT and EfficientNet have the best performances among other models. However, it is important to mention that the AttMIL method reached a very close performance to classical weakly-supervised networks for colorectal cancer (AUPRC 0.535 [0.442 - 0.632] for colorectal MSI and AUPRC 0.526 [0.480 - 0.526] for colorectal BRAF prediction). Based on the baseline value of 0.14 for colorectal MSI and 0.11 for colorectal BRAF, this performance shows the high capacity of deep learning networks in predicting colorectal MSI status directly from the WSI.

Prediction of molecular alterations in gastric and bladder cancer

While colorectal cancer is among the most widely studied tumor types in computational pathology, it is important to validate computational methods also in rarer tumor types (Echle et al., 2020). Therefore, we tested all six algorithms on the prediction of the clinically relevant alterations MSI and Epstein-Barr Virus (EBV) in gastric cancer and FGFR3 mutations in bladder cancer. We found that the overall performance in our proprietary datasets (BERN for gastric, AACHEN for bladder cancer) was lower than for colorectal cancer, which is in line with previous studies (Kather et al., 2019, Loeffler et al., 2021). The highest AUROCs were 0.761 (0.680 - 0.836) for MSI in gastric cancer (N=302 patients), 0.853 (0.645 - 0.991) for EBV in gastric cancer (N=304 patients) and 0.751 (0.667 - 0.827) for FGFR3 in bladder cancer (N=183 patients, Table 2). The highest performance was achieved by ResNet in gastric MSI and by EfficientNet in gastric EBV and by MIL in bladder FGFR3. In the external validation experiment for gastric cancer (TCGA-STAD

Table 2
Performance statistics for within-cohort experiments. Performance was assessed by stratified three-fold patient-level cross-validation. Performance is reported as patient-level area under the receiver operating curve (AUROC) with a 95% confidence interval obtained by 1000x bootstrapping. Pink = best, green = second-best. For RCC subtyping, AUROCs from top to bottom refer to clear cell, chromophobe and papillary RCC.

	Renal Cell	Colorectal	Colorectal	Gastric	Gastric	
	Ca.subtypeTCGAN= 897	MSIDACHSN=2039	BRAFDACHSN=2075	MSIBERNN=302	EBVBERNN=304	BladderFGFR3AACHENN=183
ResNet	0.976 (0.967-0.986) 0.987 (0.981-0.993) 0.980 (0.971-0.987)	0.917 (0.895-0.938)	0.845 (0.812-0.875)	0.761 (0680-0.836)	0.773 (0.567-0.965)	0.746 (0.656-0.831)
EfficientNet	0.983 (0.975-0.990) 0.992 (0.987-0.997) 0.986 (0.980-0.992)	0.930 (0.906-0.950)	0.856 (0.825-0.887)	0.754 (0.673-0.834)	0.853 (0.645-0.991)	0.736 (0.648-0.814)
ViT	0.977 (0.967-0.985) 0.984 (0.970-0.994) 0.985 (0.979-0.991)	0.906 (0.881-0.929)	0.804 (0.769-0.841)	0.732 (0.657-0.801)	0.792 (0.541-0.972)	0.729 (0.634-0.820)
MIL	0.951 (0.935-0.964) 0.955 (0.934-0.971) 0.943 (0.925-0.959)	0.709 (0.675-0.742)	0.629 (0.58-0.676)	0.512 (0.427-0.591)	0.554 (0.581-.0.809)	0.751 (0.667-0.827)
AttMIL	0.979 (0.976-0.988) 0.983 (0.972-0.992) 0.979 (0.971-0.985)	0.880 (0.851-0.909)	0.803 (0.765-0.842)	0.700 (0.618 - 0.775)	0.681 (0.502-0.851)	0.731 (0.648-0.818)
CLAM	0.969 (0.958-0.978) 0.972 (0.957-0.984) 0.973 (0.962-0.981)	0.795 (0.763-0.828)	0.671 (0.619-0.714)	0.554 (0.458-0.655)	0.801 (0.613-0.942)	0.546 (0.450-0.640)

with N=327 patients for MSI, N=327 patients for EBV), the resulting performance differences were much less clear-cut (Table 3), with no consistently best-performing method. However, for external validation of FGFR3 analysis in bladder cancer (TCGA-BLCA, N=325 patients), ViT and ResNet again outperformed all other approaches, reaching AUROCs of 0.775 (0.704 - 0.840) and 0.773 (0.699 - 0.844), respectively. The difference between ResNet and MIL and CLAM was statistically significant (p<=0.03, Suppl. Table 2).

Overall assessment of classifier performance for mutation prediction

Finally, we systematically analyzed performance differences between the six classifiers in all five mutation prediction tasks. Each method was compared to the five other methods in five tasks, yielding 25 comparisons per method. The classical weakly-supervised approach with ResNet significantly ($p<0.03$, $z>2$) outperformed other methods in 5/25 tasks and was never significantly outperformed by another method (Suppl.Table 2). ViT significantly ($p<0.03$, $z>2$) outperformed other methods in 7/25 tasks and was never significantly outperformed (Suppl. Table 4). EfficientNet outperformed other methods in 7/25 tasks and was never significantly outperformed (Suppl. Table 3). Classical MIL was outperformed by other methods in 15/25 tasks (Suppl. Table 5). AttMIL outperformed other methods in 4/25 tasks and was never outperformed by other methods (Suppl. Table 6). CLAM outperformed other methods in 1/25 mutation prediction tasks but was outperformed in 9/25 tasks (Suppl. Table 7). Overall, we conclude that classical weakly-supervised pipelines with EfficientNet and ViT-based network architectures are reasonable algorithm choices for the prediction of molecular alterations from routine histology in solid tumors. However, it should be acknowledged that the training process is much more computationally intensive and thus much slower for classical weakly supervised pipelines compared to MIL-based approaches (Suppl. Figure 8). This is partially offset by the fact that MIL-based approaches require a computationally expensive feature extraction by a pre-trained neural network before training the actual classification model.

Explainability of the performance differences

To understand the reason for the observed performance differences of the methods, we systematically compared which image tiles were assigned the highest scores by each method, in all classification tasks. We found that for renal cell carcinoma subtyping - a task in which all methods performed almost equally well - the highest scoring tiles showed plausible histopathological patterns for all classes for all methods. Consistently, tiles with high prediction scores for clear cell RCC showed carcinoma cells with clear cytoplasm; tiles predictive of chromophobe RCC showed a perinuclear halo characteristic of this subtype, and tiles with high scores for papillary RCC showed a papillary tissue architecture (Figure 3). In contrast, for MSI prediction in colorectal cancer - a task in which classical end-to-end methods outperformed MIL-based methods - the typical MSI-like morphology (Greenson et al., 2009) includes poor differentiation, mucinous differentiation, and tumor-infiltrating lymphocytes. These patterns were prominently visible in highly scoring tiles selected by high-performing meth-

Table 3
Performance statistics for external validation experiments. Performance is reported as patient-level area under the receiver operating curve (AUROC) with a 95% confidence interval obtained by 1000x bootstrapping. Pink = best, green = second-best method. For RCC subtyping, AUROCs from top to bottom refer to clear cell, chromophobe and papillary RCC. Statistical significance is reported in Suppl. Tables 2–7.

	Renal Cell Ca.subtypeAACHENN=249	Colorectal MSITCGAN=426	Colorectal BRAFTCGAN=500	Gastric MSITCGAN=327	Gastric EBVTCGAN=327	BladderFGFR3TCGAN=325
ResNet	0.971 (0.952-0.986) 0.949 (0.897-0.989) 0.980 (0.963-0.994)	0.852 (0.792-0.903)	0.777 (0.703-0.837)	0.657 (0.582-0.735)	0.779 (0.664-0.883)	0.773 (0.699-0.844)
EfficientNet	0.958 (0.928-0.982) 0.944 (0.890-0.987) 0.969 (0.930-0.995)	0.883 (0.829-0.928)	0.808 (0.749-0.861)	0.739 (0.668-0.810)	0.787 (0.675-0.887)	0.772 (0.703-0.838)
ViT	0.961 (0.933-0.982) 0.957 (0.912-0.990) 0.963 (0.926-0.992)	0.885 (0.834-0.926)	0.786 (0.719-0.844)	0.727 (0.650-0.798)	0.775 (0.661-0.870)	0.775 (0.704-0.840)
MIL	0.941 (0.910-0.967) 0.974 (0.778-0.961) 0.931 (0.900-0.960)	0.585 (0.506-0.667)	0.611 (0.531-0.687)	0.596 (0.521-0.678)	0.795 (0.712-0.872)	0.597 (0.508-0.683)
AttMIL	0.964 (0.938-0.983) 0.948 (0.862-0.994) 0.962 (0.930-0.986)	0.819 (0.756-0.874)	0.744 (0.672-0.808)	0.728 (0.659-0.798)	0.732 (0.619-0.836)	0.673 (0.586-0.757)
CLAM	0.966 (0.946-0.986) 0.927 (0.812-0.995) 0.963 (0.939-0.983)	0.656 (0.581-0732)	0.673 (0.595-0.752)	0.709 (0.642-0.780)	0.813 (0.712-0.893)	0.632 (0.545-0.713)

ods ResNet, EfficientNet and ViT. In contrast, MIL-based methods assigned the highest prediction scores to image tiles at the tissue boundary, less than half of which clearly showed MSI-like morphology (Figure 4). In addition, we analyzed the slide-level heatmaps of all model predictions for the RCC classification task (Figure 5). In this visualization, we found that MIL-based approaches generally yield a clearer outline of the actual tumor location, but the overall (slide-level) score is often better in the classical weakly-supervised approaches. This mirrors and supports the higher AUROCs for the classical approaches and motivates future studies to benchmark AUROC and visual quality of heatmaps.

Discussion

Summary of key findings

In this study, we provide a systematic benchmark for six AI algorithms applied to six clinical problems in computational pathology. The selection of these algorithms was motivated by a systematic search of the applied research literature and was representative of the research field. To benchmark these six pipelines, we selected six clinically relevant tasks which were previously addressed in one or several publications and are of direct clinical relevance. (Kather et al., 2020, Kather et al., 2019, Loeffler et al., 2021, Lu et al., 2021, Velmahos et al., 2021) As a result of this systematic benchmarking, we demonstrate that morphological subtyping of renal cell carcinoma (RCC) is an easy task in which most methods reach a high performance (Table 1 and Table 2), without significant differences between methods except for classical MIL vs. ResNet (Suppl. Table 2, 3, 4, 5, 6, and 7). By counting the number of times each method outperformed other methods (or was outperformed by other methods), a tentative ranking of the approaches in our benchmark experiments is as follows: EfficientNet and ViT (Suppl.Table 3 and 4), followed by ResNet (Suppl. Table 2), followed by AttMIL (Suppl. Table 6), followed by CLAM (Suppl. Table 7) and lastly, classical MIL (Suppl. Table 5). While the best approach varied between the different benchmark tasks in cross-validated within-cohort (Table 2) and external validation experiments (Table 3), an interesting and perhaps unexpected result is that classical weakly-supervised approaches often outperform more sophisticated MIL-based approaches, even modern attention-based MIL pipelines. Additionally Weakly-supervised approaches showed also higher AUPRC values for the external validation experiments. While AttMIL had very similar performance to the classic weakly-supervised models, it also reaches high AUPRC values in comparison to the baseline value for each dataset. In these classical approaches, all tiles inherit the slide label, which is a strong simplification. Although this simplification leads to label noise (the ground truth label is assigned to all tiles generated from a slide, not just the tumor tissue), this does not seem to impair performance when a large portion of the slide is a tumor, as in the surgical resection specimen in this study. In general, the performance of the models for gastric cancer is lower in comparison to colorectal cancer. While this finding is in line with previous studies (Echle et al., 2020, Muti et al., 2021), it is conceivably due to more diverse histological patterns in gastric cancer, which makes

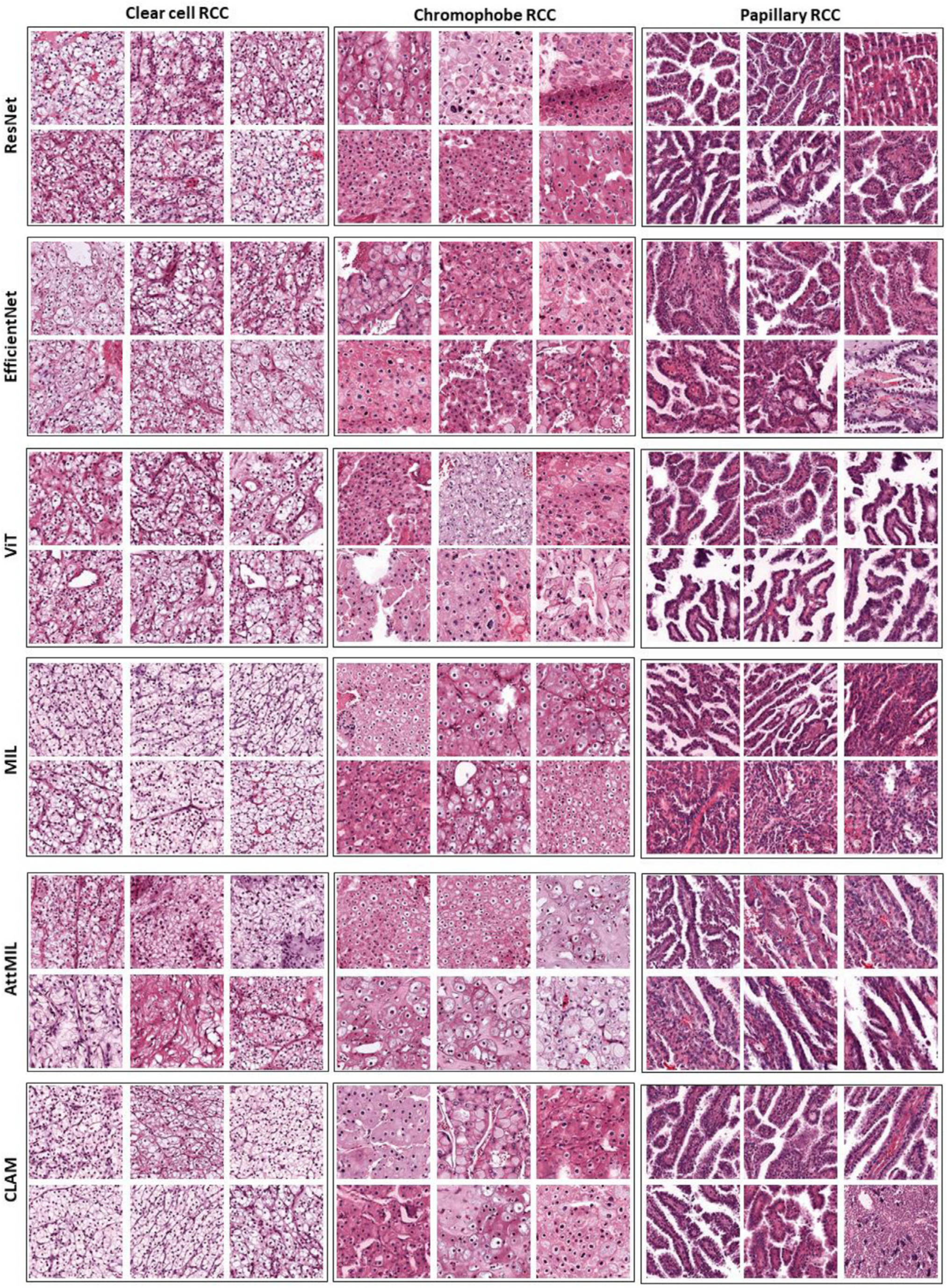

Fig. 3. Explainability of subtyping of renal cell carcinoma (RCC). The six randomly selected high-scoring tiles from 25 high-scoring tiles (5 high-score tiles per 5 high-score patients) in the external validation experiment as selected by each method are displayed. For this benchmark task, all six methods achieved a high performance. Correspondingly, all methods succeeded in selecting image tiles with patterns representative of known features of RCC subtypes.

it harder for the ANNs to detect a molecular subgroup like MSI (Suppl. Figure 9). In the next sections, we will discuss possible reasons for these observations, as well as further steps towards the development of new pipelines.

Classical weakly-supervised methods perform well despite label noise

While classical approaches, specially ResNet-based methods, have been used since 2018 (Bhatt et al., 2021, Bychkov et al., 2021, Coudray et al., 2018, Hinata and Ushiku, 2021, Kather et al., 2019), MIL has been first used in a large-scale computational pathology study in 2019 (Campanella et al., 2019). While classical MIL is susceptible to artifacts and classifier instability, the newer MIL-based variant CLAM has been shown to be more robust and powerful. (Lu et al., 2021) CLAM performs well for morphological subtyping of lung cancer and renal cell carcinoma (Lu et al., 2021) as well as for prediction of primary tumor type from metastatic tissue (Lu et al., 2021, Lu et al., 2021) Similarly, our implementation of AttMIL is a simplification of CLAM, but still conceptually superior to classical MIL. In our study, all MIL-based approaches yielded visually more appealing prediction heatmaps compared to classical weakly-supervised workflows

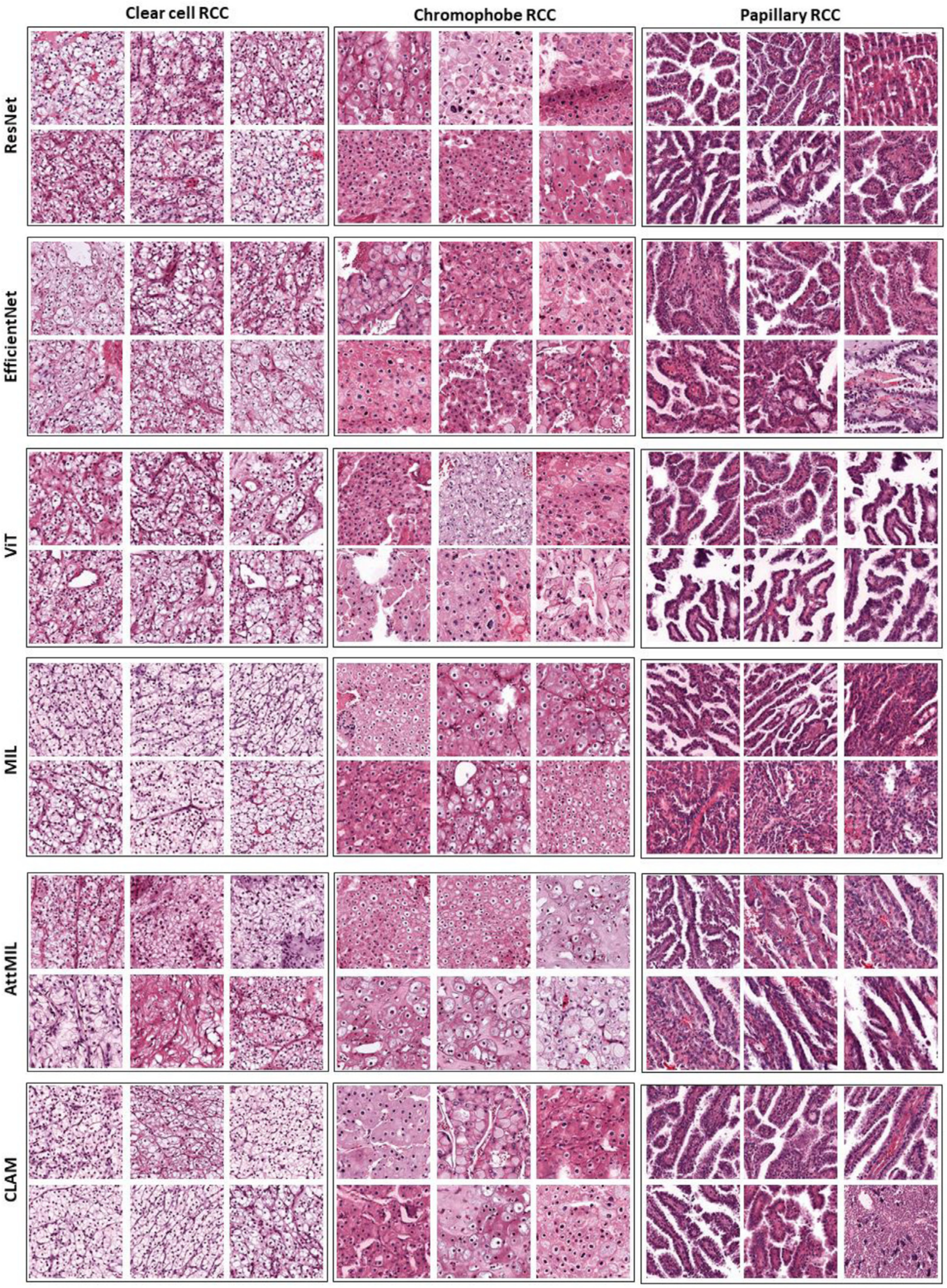

Fig. 4. Explainability of microsatellite instability (MSI) prediction in colorectal cancer (CRC). The highest-scoring tiles for the five highest-scoring patients in the external validation experiment are displayed. Resnet, EfficientNet, and ViT achieved the highest performance. This corresponds to a selection of biologically plausible tiles, showing poorly differentiated, mucinous tumors for MSI. Conversely, MIL, AttMILand CLAM selected tiles with tissue edges and other artifacts, corresponding to their lower performance.

(Figure 5). However, this did not translate into a higher performance, as the prediction scores for the true class were generally higher in classical weakly-supervised approaches, as shown for representative slides (Figure 5) and confirmed by the overall higher AUROCs in classical weakly-supervised approaches. A possibility for the lower performance of MIL/CLAM than the classical weakly-supervised approach is that an off-the-shelf pre-trained network is used for feature extraction in MIL/CLAM whereas networks are trained directly on images in classical weakly-supervised approaches. Additionally, in MIL-based approaches, all tiles are aggregated in a bag which inherits the slide label and is shown repeatedly to the network. In contrast, in classic weakly-supervised models, each tile inherits the slide label. Each slide yields 100s of tiles and therefore, the network is trained on 100s of instances of this slide with hard labels. This is an efficient form of data augmentation which is missing from MIL. Also, Because MIL-based workflows learn to focus their attention on the tumor tissue, they might miss visual clues in the normal tissue around the tumor. Previous work by Brockmoeller et al. (Brockmoeller et al., 2022) has shown that even supposedly “normal” tissue can contain

Table 4
Area under the precision recall curve (AUPRC) for all external validation experiments. Pink = best, green = second-best method.

	Renal Cell Ca. subtype AACHEN N=249	Colorectal MSI TCGA N=426	Colorectal BRAF TCGA N=500	Gastric MSI TCGA N=327	Gastric EBV TCGA N=327	Bladder FGFR3 TCGA N=325
ResNet	0.99(0.983-0.996)0.749(0.514-0.914)0.907(0.815-0.973)	0.541(0.409-0.670)	0.324(0.229-0.448)	0.285(0.205-0.413)	0.325(0.165-0.509)	0.428(0.295-0.565
EfficientNet	0.983 (0.963-0.994) 0.744 (0.546-0.899) 0.896 (0.800-0.979)	0.668 (0.548-0.775)	0.36 (0.253-0.487)	0.38 (0.265-0.531)	0.271 (0.146-0.453)	0.385 (0.254-0.53)
ViT	0.986 (0.973-0.994) 0.732 (0.484-0.896) 0.896 (0.802-0.966)	0.672 (0.558-0.769)	0.304 (0.215-0.414)	0.376 (0.258-0.530)	0.354 (0.173-0.550)	0.394 (0.275-0.548)
MIL	0.982 (0.971-0.990) 0.713 (0.514-0.883) 0.695 (0.514-0.883)	0.196 (0.138-0.283)	0.167 (0.115-0.253)	0.226 (0.164-0.337)	0.199 (0.116-0.337)	0.212 (0.137-0.318)
AttMIL	0.953 (0.932-0.974) 0.751 (0.574-0.876) 0.831 (0.740-0.907)	0.535 (0.442-0.632)	0.526 (0.480-0.569)	0.373 (0.200-0.537)	0.422 (0.308-0.528)	0.426 (0.328-0520)
CLAM	0.989 (0.981-0.996) 0.829 (0.652-0.954) 0.843 (0.738-0.931	0.221 (0.156-0.304)	0.248 (0.171-0.362)	0.304 (0.212-0.424)	0.272 (0.141-0.451)	0.267 (0.166-0.394)

information which the classical weakly supervised methods can extract.

Collectively, our findings demonstrate that researchers should not entirely rely on MIL-based approaches without testing them against a simpler and potentially more powerful classical weakly-supervised approach for their specific problem. In general, our findings motivate a future development of robust pipelines which combine the attention mechanism of modern MIL-based approaches with the overall higher performance of classical weak supervision. Some previous studies have used MIL for needle-in-a-haystack problems, such as the detection of small nests of tumor cells in biopsy tissue (Campanella et al., 2019) or in lymph nodes. (Ehteshami Bejnordi et al., 2017) Because the present study was focused on the prediction of tumor subtypes and molecular alterations, we did not include such a problem in the study. In summary, our benchmark study provides important actionable advice for future studies and real-world applications on surgical resection tissue.

Vision transformers are a new class of highly performing models

Within weakly-supervised workflows, convolutional neural networks (CNNs) are the de-facto standard architecture. Recently, Vision transformers (ViTs) have become available and shown promising performance in computer vision tasks (Dosovitskiy et al., 2020). Therefore, in the present study, we also benchmarked a new classical weakly-supervised pipeline using a ViT instead of CNNs. Our data show that ViT performed on par with but never significantly outperformed the CNNs (Suppl. Table 4). However, the ViT-based classical weakly-supervised approach outperformed MIL and CLAM. This finding is of high practical relevance for academic and commercial actors in computational pathology, as ViTs represents a relatively novel technology, which has been broadly applied outside of medicine but is still new to computational pathology.

Limitations of this study and outlook

There are multiple limitations of our study: it is in the nature of technical benchmarks that neither all possible technical approaches nor all possible applications can be evaluated. We motivate our selection of the algorithms through a systematic literature review in which we demonstrate that the selected pipelines represent the majority of the computational pathology literature in applied research studies indexed in PubMed. However, it is possible that other, less widely used approaches are even better than the ones we tested. Regarding the clinical applications, we investigate molecular subtyping in multiple tumor types. While these tasks are all similar from a technical point of view (they are all binary classification tasks), they span a range of different biomedical applications. Some of these prediction tasks are easy (classification of renal cell carcinoma). Other tasks are harder but the patterns which need to be detected are known (MSI) and some are harder and the target patterns are not fully known (FGFR3 in bladder cancer). However, future work should validate our findings in other prediction tasks, especially also in regression tasks, which are less commonly found in the applied computational pathology research literature. Importantly, as part of our study, we release an open-source workflow that includes all five approaches: the histology image analysis package (HIA). HIA is a comprehensive PyTorch-

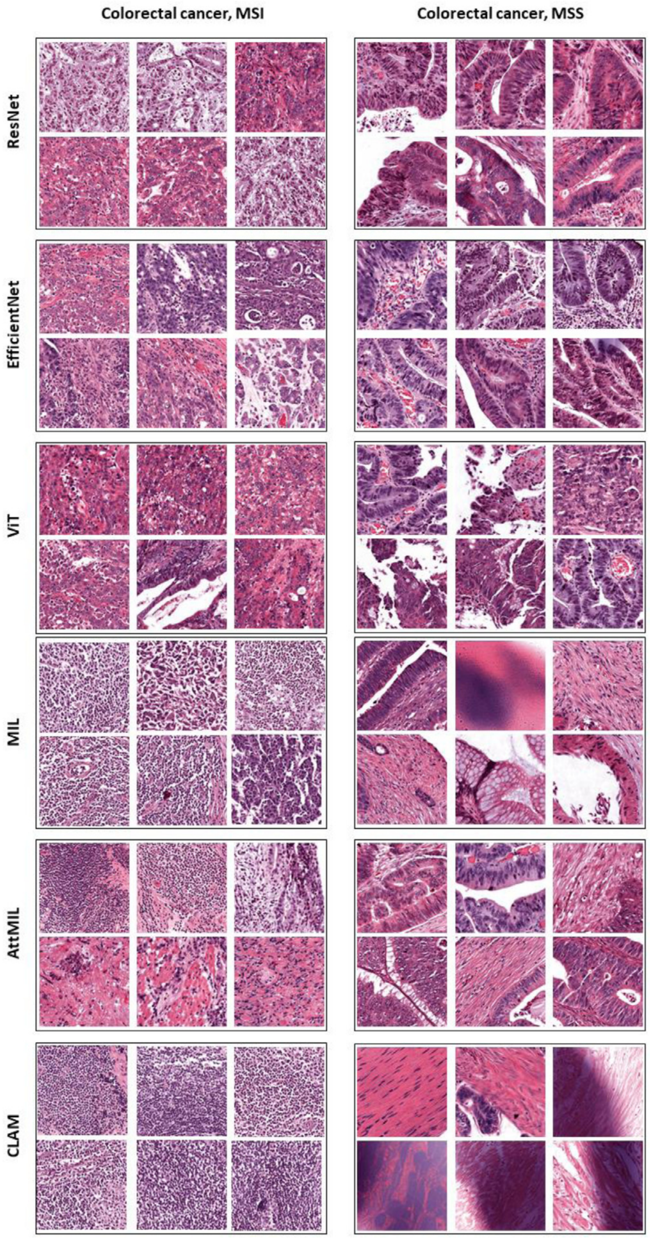

Fig. 5. Whole slide prediction maps of representative cases in the renal cell carcinoma dataset. Each column represents one case of a given class. Tile predictions are visualized as a heatmap for each of the six methods. Slide-level averages are shown next to each map.

based library that enables academic and commercial researchers to easily benchmark all tested methods on their own datasets, using just a single implementation. After initial submission of our article, multiple other open-source pipelines for computational pathology have become available, such as DeepMed (van Treeck et al., 2021), tiatoolbox (Pocock et al., 2021) and slideflow (Dolezal et al., 2021). These packages contain different implementations of many of the same algorithms used in our study, so that our findings could be helpful for users of these open-source pipelines when selecting a specific algorithm for a given image analysis problem.

Conclusion

In this study, we provide a large-scale benchmark of multiple AI approaches in computational pathology using multiple large patient cohorts. We provide HIA, an easy-to-use computational implementation which is not limited to one particular method and is therefore reusable and extensible. Surprisingly, for the prediction of molecular alterations, the classical weakly-supervised workflow was consistently superior to MIL. This provides researchers with a clear guideline and with tools of which AI methods should be used in digital pathology. In addition, for the first time, we use Vision Transformers (ViT) in computational pathology, which shows promise for future applications. Overall, our findings highlight the need to thoroughly benchmark new analysis pipelines in computational pathology against established and simpler ones.

Funding

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 #70113864). 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 NCT Tissue Bank at the Institute of Pathology, University of Heidelberg; 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). PB is supported by the DFG, German Research Foundation (Project-IDs 322900939, 454024652, 432698239, 445703531, 445703531), European Research Council (ERC; Consolidator Grant AIM.imaging.CKD, No 101001791), Federal Ministry of Education and Research(STOP-FSGS-01GM1901A), and Federal Ministry of Economic Affairs and Energy (EMPAIA, No. 01MK2002A).NTG is funded by the DFG (GA 1384/3-1, GA 1384/5-1).

Declaration of competing interest

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, Handschuhsheimer Landstr. 9/1, 69120 Heidelberg). No other potential conflicts of interest are reported by any of the authors.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.media.2022.102474.

References

Alwers, E, Bläker, H, Walter, V, Jansen, L, Kloor, M, Arnold, A, et al., 2019. External validation of molecular subtype classifications of colorectal cancer based on microsatellite instability, CIMP, BRAF and KRAS. BMC Cancer 19, 681.
Bengs, M, Bockmayr, M, Schüller, U, Schlaefer, A., 2021. Medulloblastoma tumor classification using deep transfer learning with multi-scale EfficientNets. Medical Imaging 2021: Digital Pathology. International Society for Optics and Photonics.
Berrada L, Zisserman A, Pawan Kumar M. Smooth Loss Functions for Deep Top-k Classification. arXiv [cs.LG]. 2018. Available: http://arxiv.org/abs/1802.07595.
Bhatt, AR, Ganatra, A, Kotecha, K., 2021. Cervical cancer detection in pap smear whole slide images using convNet with transfer learning and progressive resizing. PeerJ Comput Sci 7, e348.
Bilal M, Raza SEA, Azam A, Graham S, Ilyas M. Development and validation of a weakly supervised deep learning framework to predict the status of molecular pathways and key mutations in colorectal …. The Lancet Digital. 2021. Available: https://www.sciencedirect.com/science/article/pii/S2589750021001801.
Brenner, H, Chang-Claude, J, Seiler, CM, Stürmer, T, Hoffmeister, M., 2006. Does a negative screening colonoscopy ever need to be repeated? Gut 55, 1145–1150.
Brenner, H, Chang-Claude, J, Seiler, CM, Rickert, A, Hoffmeister, M., 2011. Protection from colorectal cancer after colonoscopy: a population-based, case-control study. Ann Intern Med 154, 22–30.

Brockmoeller, S, Echle, A, Ghaffari Laleh, N, Eiholm, S, Malmstrøm, ML, Plato Kuhlmann, T, et al., 2022. Deep learning identifies inflamed fat as a risk factor for lymph node metastasis in early colorectal cancer. J Pathol 256, 269–281.

Bulten, W, Pinckaers, H, van Boven, H, Vink, R, de Bel, T, van Ginneken, B, et al., 2020. Automated deep-learning system for Gleason grading of prostate cancer using biopsies: a diagnostic study. Lancet Oncol 21, 233–241.

Bychkov, D, Linder, N, Tiulpin, A, Kücükel, H, Lundin, M, Nordling, S, et al., 2021. Deep learning identifies morphological features in breast cancer predictive of cancer ERBB2 status and trastuzumab treatment efficacy. Sci Rep 11, 4037.

Campanella, G, Hanna, MG, Geneslaw, L, Miraflor, A, Werneck Krauss Silva, V, Busam, KJ, et al., 2019. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. Nat Med 25, 1301–1309.

Cancer Genome Atlas Research Network, 2014. Comprehensive molecular characterization of gastric adenocarcinoma. Nature 513, 202–209.

Cancer Genome Atlas Research Network, 2014. Comprehensive molecular characterization of urothelial bladder carcinoma. Nature 507, 315–322.

Chen, C-L, Chen, C-C, Yu, W-H, Chen, S-H, Chang, Y-C, Hsu, T-I, et al., 2021. An annotation-free whole-slide training approach to pathological classification of lung cancer types using deep learning. Nat Commun 12, 1193.

Coudray, N, Tsirigos, A., 2020. Deep learning links histology, molecular signatures and prognosis in cancer. Nature Cancer doi:10.1038/s43018-020-0099-2.

Coudray, N, Ocampo, PS, Sakellaropoulos, T, Narula, N, Snuderl, M, Fenyö, D, et al., 2018. Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning. Nat Med 24, 1559–1567.

Das, K, Conjeti, S, Roy, AG, Chatterjee, J, Sheet, D., 2018. Multiple instance learning of deep convolutional neural networks for breast histopathology whole slide classification. In: 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018). ieeexplore.ieee.org, pp. 578–581.

DeLong, ER, DeLong, DM, Clarke-Pearson, DL., 1988. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics 44, 837–845.

Dietterich, TG, Lathrop, RH, Lozano-Pérez, T., 1997. Solving the multiple instance problem with axis-parallel rectangles. Artif Intell 89, 31–71.

Dislich, B, Blaser, N, Berger, MD, Gloor, B, Langer, R., 2020. Preservation of Epstein–Barr virus status and mismatch repair protein status along the metastatic course of gastric cancer. Histopathology 76, 740–747.

Dolezal J, Kochanny S, Howard F. jamesdolezal/slideflow: Slideflow 1.0 - Official Public Release. 2021. doi:10.5281/zenodo.5708490

Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, et al. An Image is Worth 16×16 Words: Transformers for Image Recognition at Scale. arXiv [cs.CV]. 2020. Available: http://arxiv.org/abs/2010.11929.

Echle, A, Rindtorff, NT, Brinker, TJ, Luedde, T, Pearson, AT, Kather, JN., 2020. Deep learning in cancer pathology: a new generation of clinical biomarkers. Br J Cancer 1–11.

Echle, A, Grabsch, HI, Quirke, P, van den Brandt, PA, West, NP, Hutchins, GGA, et al., 2020. Clinical-Grade Detection of Microsatellite Instability in Colorectal Tumors by Deep Learning. Gastroenterology 159, 1406–1416 e11.

Echle, A, Laleh, NG, Schrammen, PL, West, NP, Trautwein, C, Brinker, TJ, et al., 2021. Deep learning for the detection of microsatellite instability from histology images in colorectal cancer: A systematic literature review. ImmunoInformatics 3-4, 100008.

Ehteshami Bejnordi, B, Veta, M, Johannes van Diest, P, van Ginneken, B, Karssemeijer, N, Litjens, G, et al., 2017. Diagnostic Assessment of Deep Learning Algorithms for Detection of Lymph Node Metastases in Women With Breast Cancer. JAMA 318, 2199–2210.

Fu, Y, Jung, AW, Torne, RV, Gonzalez, S, Vöhringer, H, Shmatko, A, et al., 2020. Pan–cancer computational histopathology reveals mutations, tumor composition and prognosis. Nature Cancer 1, 800–810.

Gheisari, S, Catchpoole, DR, Charlton, A, Kennedy, PJ, 2018. Convolutional Deep Belief Network with Feature Encoding for Classification of Neuroblastoma Histological Images. J Pathol Inform 9, 17.

Greenson, JK, Huang, S-C, Herron, C, Moreno, V, Bonner, JD, Tomsho, LP, et al., 2009. Pathologic predictors of microsatellite instability in colorectal cancer. Am J Surg Pathol 33, 126–133.

He, K, Zhang, X, Ren, S, Sun, J, 2016. Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770–778.

Hinata, M, Ushiku, T., 2021. Detecting immunotherapy-sensitive subtype in gastric cancer using histologic image-based deep learning. Sci Rep 11, 22636.

Hoffmeister, M, Bläker, H, Jansen, L, Alwers, E, Amitay, EL, Carr, PR, et al., 2020. Colonoscopy and Reduction of Colorectal Cancer Risk by Molecular Tumor Subtypes: A Population-Based Case-Control Study. Am J Gastroenterol 115, 2007–2016.

Hurst, CD, Zuiverloon, TCM, Hafner, C, Zwarthoff, EC, Knowles, MA., 2009. A SNaPshot assay for the rapid and simple detection of four common hotspot codon mutations in the PIK3CA gene. BMC Res Notes 2, 66.

Ilse M, Tomczak JM, Welling M. Attention-based deep multiple instance learning. arXiv [cs.LG]. 2018. Available: http://proceedings.mlr.press/v80/ilse18a/ilse18a.pdf.

Jang, H-J, Song, I-H, Lee, S-H., 2021. Deep Learning for Automatic Subclassification of Gastric Carcinoma Using Whole-Slide Histopathology Images. Cancers 13. doi:10.3390/cancers13153811.

Jia, M, Jansen, L, Walter, V, Tagscherer, K, Roth, W, Herpel, E, et al., 2016. No association of CpG island methylator phenotype and colorectal cancer survival: population-based study. Br J Cancer 115, 1359–1366.

Kacew, AJ, Strohbehn, GW, Saulsberry, L, Laiteerapong, N, Cipriani, NA, Kather, JN, et al., 2021. Artificial Intelligence Can Cut Costs While Maintaining Accuracy in Colorectal Cancer Genotyping. Frontiers in Oncology doi:10.3389/fonc.2021.630953.

Kanavati, F, Toyokawa, G, Momosaki, S, Rambeau, M, Kozuma, Y, Shoji, F, et al., 2020. Weakly-supervised learning for lung carcinoma classification using deep learning. Sci Rep 10, 9297.

Kather, JN, Calderaro, J., 2020. Development of AI-based pathology biomarkers in gastrointestinal and liver cancer. Nat Rev Gastroenterol Hepatol doi:10.1038/s41575-020-0343-3.

Kather, JN, Pearson, AT, Halama, N, Jäger, D, Krause, J, Loosen, SH, et al., 2019. Deep learning can predict microsatellite instability directly from histology in gastrointestinal cancer. Nat Med 25, 1054–1056.

Kather, JN, Heij, LR, Grabsch, HI, Loeffler, C, Echle, A, Muti, HS, et al., 2020. Pan–cancer image-based detection of clinically actionable genetic alterations. Nat Cancer 1, 789–799.

Kopetz, S, Grothey, A, Yaeger, R, Van Cutsem, E, Desai, J, Yoshino, T, et al., 2019. Encorafenib, Binimetinib, and Cetuximab in BRAF V600E-Mutated Colorectal Cancer. N Engl J Med 381, 1632–1643.

Kosaraju, SC, Hao, J, Koh, HM, Kang, M., 2020. Deep-Hipo: Multi-scale receptive field deep learning for histopathological image analysis. Methods 179, 3–13.

Kott, O, Linsley, D, Amin, A, Karagounis, A, Jeffers, C, Golijanin, D, et al., 2021. Development of a Deep Learning Algorithm for the Histopathologic Diagnosis and Gleason Grading of Prostate Cancer Biopsies: A Pilot Study. Eur Urol Focus 7, 347–351.

Li J, Chen W, Huang X, Hu Z, Duan Q, Li H, et al. Hybrid Supervision Learning for Pathology Whole Slide Image Classification. arXiv [cs.CV]. 2021. Available: http://arxiv.org/abs/2107.00934.

Liu, Y, Sethi, NS, Hinoue, T, Schneider, BG, Cherniack, AD, Sanchez-Vega, F, et al., 2018. Comparative Molecular Analysis of Gastrointestinal Adenocarcinomas. Cancer Cell 33, 721–735 e8.

Loeffler, CML, Bruechle, NO, Jung, M, Seillier, L, Rose, M, Laleh, NG, et al., 2021. Artificial Intelligence–based Detection of FGFR3 Mutational Status Directly from Routine Histology in Bladder Cancer: A Possible Preselection for Molecular Testing? European Urology Focus doi:10.1016/j.euf.2021.04.007.

Lu, MY, Williamson, DFK, Chen, TY, Chen, RJ, Barbieri, M, Mahmood, F., 2021. Data-efficient and weakly supervised computational pathology on whole-slide images. Nature Biomedical Engineering 1–16.

Lu, MY, Chen, TY, Williamson, DFK, Zhao, M, Shady, M, Lipkova, J, et al., 2021. AI-based pathology predicts origins for cancers of unknown primary. Nature doi:10.1038/s41586-021-03512-4.

Macenko, M, Niethammer, M, Marron, JS, Borland, D, Woosley, JT, Guan, Xiaojun, et al., 2009. A method for normalizing histology slides for quantitative analysis. In: 2009 IEEE International Symposium on Biomedical Imaging: From Nano to Macro, pp. 1107–1110.

Molecular testing strategies for Lynch syndrome in people with colorectal cancer - NICE Guidance. [cited 13 Nov 2019]. Available: https://www.nice.org.uk/guidance/dg27/chapter/1-Recommendations.

Muti HS, Loeffler C, Echle A, Heij LR, Buelow RD. The Aachen protocol for deep learning histopathology: a hands-on guide for data preprocessing. 2020. Available: https://scholar.archive.org/work/5txzjhu6tjgmvg4cyxi3tendpi/access/wayback/https://zenodo.org/record/3694994/files/Aachen%20Protocol%20for%20Deep%20Learning%20Histopathology%20v0.2.pdf.

Muti HS, Heij LR, Keller G, Kohlruss M, Langer R, Dislich B, et al. Development and validation of deep learning classifiers to detect Epstein-Barr virus and microsatellite instability status in gastric cancer: a retrospective multicentre cohort study. The Lancet Digital Health. 2021;0. doi:10.1016/S2589-7500(21)00133-3

Pinckaers, H, Bulten, W, van der Laak, J, Litjens, G., 2021. Detection of Prostate Cancer in Whole-Slide Images Through End-to-End Training With Image-Level Labels. IEEE Trans Med Imaging 40, 1817–1826.

Pocock J, Graham S, Vu QD, Jahanifar M, Deshpande S, Hadjigeorghiou G, et al. TIA-Toolbox: An End-to-End Toolbox for Advanced Tissue Image Analytics. bioRxiv. 2021. p. 2021.12.23.474029. doi:10.1101/2021.12.23.474029.

Ricketts, CJ, De Cubas, AA, Fan, H, Smith, CC, Lang, M, Reznik, E, et al., 2018. The Cancer Genome Atlas Comprehensive Molecular Characterization of Renal Cell Carcinoma. Cell Rep 23, 313–326 e5.

Rony J, Belharbi S, Dolz J, Ben Ayed I, McCaffrey L, Granger E. Deep weakly-supervised learning methods for classification and localization in histology images: a survey. arXiv [cs.CV]. 2019. Available: http://arxiv.org/abs/1909.03354.

Saillard, C, Schmauch, B, Laifa, O, Moarii, M, Toldo, S, Zaslavskiy, M, et al., 2020. Predicting survival after hepatocellular carcinoma resection using deep-learning on histological slides. Hepatology doi:10.1002/hep.31207.

Schirris Y, Gavves E, Nederlof I, Horlings HM, Teuwen J. DeepSMILE: Self-supervised heterogeneity-aware multiple instance learning for DNA damage response defect classification directly from H&E whole-slide images. arXiv [eess.IV]. 2021. Available: http://arxiv.org/abs/2107.09405.

Schrammen, PL, Ghaffari Laleh, N, Echle, A, Truhn, D, Schulz, V, Brinker, TJ, et al., 2021. Weakly supervised annotation-free cancer detection and prediction of genotype in routine histopathology. J Pathol doi:10.1002/path.5800.

Sha, L, Osinski, BL, Ho, IY, Tan, TL, Willis, C, Weiss, H, et al., 2019. Multi-Field-of-View Deep Learning Model Predicts Nonsmall Cell Lung Cancer Programmed Death-Ligand 1 Status from Whole-Slide Hematoxylin and Eosin Images. J Pathol Inform 10, 24.

Shaban, M, Awan, R, Fraz, MM, Azam, A, Tsang, Y-W, Snead, D, et al., 2020. Context-Aware Convolutional Neural Network for Grading of Colorectal Cancer Histology Images. IEEE Trans Med Imaging 39, 2395–2405.

Sharma, Y, Shrivastava, A, Ehsan, L, Moskaluk, CA, Syed, S, Brown, D., 2021. Cluster-to-Conquer: A Framework for End-to-End Multi-Instance Learning for Whole Slide Image Classification. In: Proceedings of the Fourth Conference on Medical Imaging with Deep Learning, pp. 682–698 PMLR; 07–09 Jul.

Skrede, O-J, De Raedt, S, Kleppe, A, Hveem, TS, Liestøl, K, Maddison, J, et al., 2020. Deep learning for prediction of colorectal cancer outcome: a discovery and validation study. Lancet 395, 350–360.

Ström, P, Kartasalo, K, Olsson, H, Solorzano, L, Delahunt, B, Berney, DM, et al., 2020. Artificial intelligence for diagnosis and grading of prostate cancer in biopsies: a population-based, diagnostic study. Lancet Oncol 21, 222–232.

Sudharshan, PJ, Petitjean, C, Spanhol, F, Oliveira, LE, Heutte, L, Honeine, P., 2019. Multiple instance learning for histopathological breast cancer image classification. Expert Syst Appl 117, 103–111.

Tan, M, Le, Q., 2019. Efficientnet: Rethinking model scaling for convolutional neural networks. In: International Conference on Machine Learning. PMLR, pp. 6105–6114.

Tan M, Le QV. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. arXiv [cs.LG]. 2019. Available: http://arxiv.org/abs/1905.11946.

Tan M, Le QV. EfficientNetV2: Smaller Models and Faster Training. arXiv [cs.CV]. 2021. Available: http://arxiv.org/abs/2104.00298.

Touvron H, Cord M, Douze M, Massa F, Sablayrolles A, Jégou H. Training data-efficient image transformers & distillation through attention. arXiv [cs.CV]. 2020. Available: http://arxiv.org/abs/2012.12877.

van Treeck M, Cifci D, Laleh NG, Saldanha OL, Loeffler CML, Hewitt KJ, et al. DeepMed: A unified, modular pipeline for end-to-end deep learning in computational pathology. bioRxiv. 2021. p. 2021.12.19.473344. doi:10.1101/2021.12.19.473344.

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention Is All You Need. arXiv [cs.CL]. 2017. Available: http://arxiv.org/abs/1706.03762.

Velmahos, CS, Badgeley, M, Lo, Y-C., 2021. Using deep learning to identify bladder cancers with FGFR-activating mutations from histology images. Cancer Med 10, 4805–4813.

Wang, X, Chen, H, Gan, C, Lin, H, Dou, Q, Tsougenis, E, et al., 2020. Weakly Supervised Deep Learning for Whole Slide Lung Cancer Image Analysis. IEEE Trans Cybern 50, 3950–3962.

Wang, X, Zou, C, Zhang, Y, Li, X, Wang, C, Ke, F, et al., 2021. Prediction of BRCA Gene Mutation in Breast Cancer Based on Deep Learning and Histopathology Images. Front Genet 12, 661109.

Xu, Y, Mo, T, Feng, Q, Zhong, P, Lai, M, Chang, EI, 2014. Deep learning of feature representation with multiple instance learning for medical image analysis. In: 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). ieeexplore.ieee.org, pp. 1626–1630.

Yamamoto, Y, Tsuzuki, T, Akatsuka, J, Ueki, M, Morikawa, H, Numata, Y, et al., 2019. Automated acquisition of explainable knowledge from unannotated histopathology images. Nat Commun 10, 5642.

Yamashita, R, Long, J, Longacre, T, Peng, L, Berry, G, Martin, B, et al., 2021. Deep learning model for the prediction of microsatellite instability in colorectal cancer: a diagnostic study. Lancet Oncol 22, 132–141.

Yao, J, Zhu, X, Jonnagaddala, J, Hawkins, N, Huang, J., 2020. Whole slide images based cancer survival prediction using attention guided deep multiple instance learning networks. Med Image Anal 65, 101789.

Zhu, M, Ren, B, Richards, R, Suriawinata, M, Tomita, N, Hassanpour, S., 2021. Development and evaluation of a deep neural network for histologic classification of renal cell carcinoma on biopsy and surgical resection slides. Sci Rep 11, 7080.