Clinical actionability of triaging DNA mismatch repair deficient colorectal cancer from biopsy samples using deep learning

Wu Jiang,[a,b,1] Wei-Jian Mei,[a,b,1] Shuo-Yu Xu,[c,d,1] Yi-Hong Ling,[a,e,1] Wei-Rong Li,[f] Jin-Bo Kuang,[c] Hao-Sen Li,[c] Hui Hui,[c] Ji-Bin Li,[a,g] Mu-Yan Cai,[a,e] Zhi-Zhong Pan,[a,b] Hui-Zhong Zhang,[a,e]*** Li Li,[a,h]** and Pei-Rong Ding[a,b]*

[a]Sun Yat-sen University Cancer Center; State Key Laboratory of Oncology in South China; Collaborative Innovation Center for Cancer Medicine, Guangzhou, 510060, PR China
[b]Department of Colorectal Surgery, Sun Yat-sen University Cancer Center, Guangzhou, PR China
[c]Bio-totem Pte Ltd, Foshan, PR China
[d]Department of General Surgery, Nanfang Hospital, Southern Medical University, Guangzhou, PR China
[e]Department of Pathology, Sun Yat-sen University Cancer Center, Guangzhou, PR China
[f]Department of General Surgery, Guangzhou First People's Hospital, Guangzhou, PR China
[g]Department of Clinical Research, Sun Yat-sen University Cancer Center, Guangzhou, PR China
[h]Department of Medical Imaging, Sun Yat-sen University Cancer Center, Guangzhou, PR China

eBioMedicine 2022;81: 104120
Published online xxx
https://doi.org/10.1016/j.ebiom.2022.104120

Summary

Background We aimed to develop a deep learning (DL) model to predict DNA mismatch repair (MMR) status in colorectal cancers (CRC) based on hematoxylin and eosin-stained whole-slide images (WSIs) and assess its clinical applicability.

Methods The DL model was developed and validated through three-fold cross validation using 441 WSIs from the Cancer Genome Atlas (TCGA) and externally validated using 78 WSIs from the Pathology AI Platform (PAIP), and 355 WSIs from surgical specimens and 341 WSIs from biopsy specimens of the Sun Yet-sun University Cancer Center (SYSUCC). Domain adaption and multiple instance learning (MIL) techniques were adopted for model development. The performance of the models was evaluated using the area under the receiver operating characteristic curve (AUROC). A dual-threshold strategy was also built from the surgical cohorts and validated in the biopsy cohort. Sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), F1-score, and the percentage of patients avoiding IHC testing were evaluated.

Findings The MIL model achieved an AUROC of 0·8888±0·0357 in the TCGA-validation cohort, 0·8806±0·0232 in the PAIP cohort, 0·8457±0·0233 in the SYSUCC-surgical cohort, and 0·7679±0·0342 in the SYSUCC-biopsy cohort. A dual-threshold triage strategy was used to rule-in and rule-out dMMR patients with remaining uncertain patients recommended for further IHC testing, which kept sensitivity higher than 90% and specificity higher than 95% on deficient MMR patient triage from both the surgical and biopsy specimens, result in more than half of patients avoiding IHC based MMR testing.

Interpretation A DL-based method that could directly predict CRC MMR status from WSIs was successfully developed, and a dual-threshold triage strategy was established to minimize the number of patients for further IHC testing.

Funding The study was funded by the National Natural Science Foundation of China (82073159, 81871971 and 81700576), the Natural Science Foundation of Guangdong Province (No. 2021A1515011792 and No.2022A1515012403) and Medical Scientific Research Foundation of Guangdong Province of China (No. A2020392).

*Corresponding author at: Department of Colorectal Surgery, Sun Yat-sen University Cancer Center, 651 Dongfeng East Road, Guangzhou 510060, PR China.
**Corresponding author at: Department of Medical Imaging, Sun Yat-sen University Cancer Center, 651 Dongfeng East Road, Guangzhou 510060, PR China.
***Corresponding author at: Department of Pathology, Sun Yat-sen University Cancer Center, 651 Dongfeng East Road, Guangzhou 510060, PR China.
E-mail addresses: zhanghuizh@sysucc.org.cn (H.-Z. Zhang), lil@sysucc.org.cn (L. Li), dingpr@sysucc.org.cn (P.-R. Ding).

[1] Wu Jiang, Wei-Jian Mei, Shuo-Yu Xu and Yi-Hong Ling contributed equally to this work.

Keywords: Colorectal cancer; Mismatch repair-deficient; Deep learning; Screening strategy; Dual-threshold

Research in context

Evidence before this study

Literature searches were conducted using PubMed and the following search terms: "colorectal cancer" AND "deep learning" OR "artificial intelligence", with no restrictions on language or publication date. Our search identified several studies that utilized deep learning to predict the MMR status of colorectal cancer (CRC) tissues directly from routine histopathology, mainly focusing on the surgical specimens. On the other hand, we found one study validated on biopsy samples using a western cohort but it only reported the association between the MSI classifier, without a clear threshold, and clinical applicability. It still requires further evidence to validate the histopathological image based MSI classifier on biopsy samples and to better demonstrate its clinical utility.

Added value of this study

In this study, we successfully developed a deep learning (DL) model to triage the MMR status of colorectal cancer from surgical specimens, which was validated on an Eastern population cohort of biopsy specimens. The DL model performed better than previously published methods, especially on biopsy samples. We also implemented a new dual-threshold triage strategy to explore the rule-in and rule-out of MMR deficient patients and recommended for further testing for the remaining patients. While maintaining a reasonable predictive performance, the proposed DL model considered that over half of the CRC patients did not need further IHC/PCR-based MMR testing.

Implications of all the available evidence

Our strategy could supplement current screening workflow by excluding over half of the biopsy specimens for further testing while maintaining a high sensitivity and specificity of MMR detection. The convenience and low cost of the proposed deep learning model could be useful in primary care hospitals and developing countries without experienced pathologists.

Introduction

Colorectal cancer (CRC) is a group of heterogenous diseases and the second most common cause of cancer death worldwide.[1,2] One of the clinically relevant subtypes of CRC is DNA mismatch repair deficient (dMMR) CRC. Patients with dMMR CRC has large numbers of mutations in repetitive DNA sequences, also called microsatellite instability (MSI).[3] dMMR can result in Lynch syndrome, a condition whereby one allele is mutated in the germline and a second mutation occurs spontaneously and sporadic cases when one allele is spontaneously mutated and the second is epigenetically silenced.[4] Patients with dMMR CRC usually respond poorly to fluorouracil-based chemotherapies but may respond to immunotherapy.[5,6] Therefore, identifying patient's MMR or MSI status is essential for treatment decision making.[7,8] In addition, patients with dMMR CRC are candidates for Lynch syndrome, who need further genetic testing. Currently, the National Comprehensive Cancer Network (NCCN) guidelines recommend universal MMR or MSI testing in all CRC patients.[9] However, these genetic or immunohistochemical (IHC) examinations are costly and time-consuming. Further, experienced pathologists and advanced laboratory equipment are needed to for accurate results, making its implementation in low-tier hospitals and developing countries very challenging.[10] On the other hand, the incidence of dMMR is only around 10−15% in CRC patients, which makes the current universal testing strategy even less cost-effective.[11] Hence, it is necessary to develop a convenient and accurate testing method with lower costs and smaller workloads to assist the selection of patients for MMR or MSI testing.

Histopathological assessment of biopsies and surgical specimens using hematoxylin and eosin (H&E) staining is a common procedure in the clinical practice. Pathologists have found visual morphological features from H&E-stained colon tissue slides, such as Crohn's-like reaction, to be associated with MMR status.[12] However, pathologists still find it challenging to accurately identify MMR status solely based on visual inspections of tissue morphology.[13] Recent advances in artificial intelligent image analysis techniques, especially deep learning approaches, have shown promising performance in various histopathological analytic tasks, including diagnosis, prognosis estimation, and gene mutation prediction.[14−17] There are growing evidences supporting the possible use of deep learning (DL) for H&E stained image-based MMR status detection in CRC, with an area-under-ROC curves (AUROC) between 0·77 and 0·96.[18−25] Thus, DL is a promising technology that could be further improve the increase detection accuracy.

In the current clinical practice, the detection of MMR status from biopsy specimens has more significant clinical value than surgical specimen because prediction from biopsy specimen is more challenging considering inadequate tumor areas and sampling site issues, and evidence of MMR detection from biopsy specimens is limited. Echle et al. found that the MMR detection performance from biopsy specimen was worse than surgical specimens but they only reported AUROC without a clinical threshold.[18]

In this study, we developed a deep learning-based system for predicting MMR status in CRCs on H&E stained whole-slide images (WSIs) utilizing domain adaption and multiple instance learning (MIL) techniques for better model generalization between surgical specimens and biopsy specimens. Dual-threshold strategy for patient triage was also used to minimize the number of patients for further IHC and molecular testing of MMR status.

Method

Patient cohorts

We retrospectively and randomly recruited 696 patients over 18 years old with pathologically confirmed colorectal cancer and IHC confirmed MMR status between January 2011 and December 2018 at the Sun Yet-sun University Cancer Center (SYSUCC, Guangzhou, China), of whom 355 underwent surgery with curative intent and 341 underwent colonoscopic examination (Supplementary Figure S3). Paraffine-embedded biopsy or gross specimens were collected and stained with H&E following routine protocols. The MMR status of each patient was determined according to the IHC assessment results of MLH1, MSH2, MSH6, and PMS2 stained slides as the NCCN guideline recommended. The absence of expression of one or more of the four DNA MMR proteins is often reported as dMMR. The MMR status of each specimen was independently confirmed by two experienced pathologists from SYSUCC. Additional information such as age at diagnosis, gender, anatomical location, pathological type, AJCC stage (Eighth Edition) and family history of cancer were collected from electronic medical records. All the H&E-stained slides were digitized at the resolution of 0·25 µm/pixel (Aperio, ScanScope AT2, Leica). Image tiles extracted from 411 colon adenocarcinoma whole-slide images (WSIs) of The Cancer Genome Atlas (TCGA)[26] and 78 WSIs of the Pathology AI Platform (PAIP)[27] were also included. The study pathologists were blinded to all clinical information and index test results and the performers of the index test were blinded to all clinical information and reference standard results. This study was performed following the STARD guidelines for Reporting Diagnostic Accuracy Studies.

Tissue classifier

The WSIs were first classified into 9 tissue types, namely background, adipose, debris/necrosis, aggregated lymphocytes, mucus, smooth muscle, normal colon mucosa, tumor-associated stroma and adenocarcinoma epithelium (Figure 1a). We used a publicly available training dataset[28] which was widely adopted for such colon tissue classification task. The original training set (NCT-CRC-HE-100K) included 100,000 non-overlapping image tiles extracted from WSIs of H&E stained tissue samples. We purposely selected the no normalization version of the dataset to avoid the stain normalization procedure, which is time-consuming and increases the method's throughput. Three-fold cross-validation was performed for training and validation. Another publicly available dataset (CRC-VAL-HE-7K)[28] was used as the external test set (7180 tiles).

The base model of the tissue classifier was Densenet121 integrated with focal loss, which has demonstrated superior performance of multi-tissue partitioning for WSI of colon samples.[29] To further enhance the generalization capacity of the tissue classifier, we incorporated a domain generalization technique (IBN-net)[30] by wrapping both batch normalization (BN) and instance normalization (IN) into deep networks. It is known that BN is important to preserve content-related information, while IN could help learning features invariant to color appearance variations. The original implementation of IBN-net was applied to Resnet, and we adapted it to Densenet following the principle of using BN and IN in the shallow layers and BN only in the deeper layers of the network (Densenet-IBN). More specifically, BN was replaced by IN after the first convolution layer, and in the transition blocks after the first and second dense block (Supplementary Figure S1). The remaining parts of the original Densenet were kept unchanged. The training parameters were kept identical for both Densenet and Densenet-IBN. Stain color augmentation was performed by shifting hue and saturation channels within the range of ±25% for both models. Other image augmentations such as flipping, rotation, and random erasing were also performed on the fly. We used focal loss as the loss function, a learning rate starting from 10-3 with weight decay of 10-5, Adam optimizer and batch size of 128. The model was trained for 40 epochs with early stopping.

Without stain normalization, the Densenet model achieved an overall good performance for all the classes in the training set (F1 = 0·984±0·006) and validation set (F1 = 0·978±0·005) but performed poorly in the test set (F1 = 0·527±0·025), indicating the weakness of generalization capacity for staining variation. On the other hand, Densenet-IBN performed well on all of the training (F1 = 0·992±0·002), validation (F1 = 0·980±0·001) and test sets (F1 = 0·682±0·049). As the primary focus of the tissue classifier was to identify tumor patches for following process, we found superior F1 scores in the tumor class of Densenet-IBN (train: 0·993±0·003; validation: 0·979±0·002; test: 0·920±0·026) than of Densenet (train: 0·984±0·007; validation: 0·978±0·005; test: 0·839±0·063). F1 scores of each tissue class and confusion matrix are summarized in Supplementary Tables S1 and S2 and Supplementary Figure S2.

MSI classifier

The training image tiles for the MSI classifier were provided by Kather et al., which was extracted from the TCGA database with WSI level slide information available for each tile.[26] Only non-normalized tumor tiles were included at the size of 512*512 pixels and resolution of 0·5 um/pixel. Three-fold cross-validation was performed at the WSI level while all the tiles from one WSI were ensured either in the training or validation set for each fold. The method to define MSI/MSS status was described in the previous publication.[31] We collected the same MSI/MSS status from it for training the MSI classifier in this study.

Fully supervised training. In fully supervised training, each tile was assigned the same label (MSI/MSS status) as the slide label it belongs to. The Densenet-IBN model was used as the deep network to train the MSI classifier at the tile level (Figure 1a). Intensive image augmentation, including flipping, Gaussian blurring, and color variation in the hue channel, was performed. Since the number of MSI and MSS tiles in the training set was unbalanced, tile re-sampling was performed to increase the MSI tiles. The model was trained for 25 epochs using cross-entropy loss function, Adam optimizer, and the learning rate of 10-5 with weight decay of 10-5 and batch size of 64.

Weakly supervised training. Although the supervised training approach was successfully adopted in several published studies[18,20,31] to classify MSI status from H&E images, it concerns assigning all tiles the same slide label on each slide as one slide might contain both MSI and MSS tiles. Hence, we treated this MSI prediction problem as a weakly supervised task and used a multiple instance learning (MIL) approach[32] to train the MSI classifier. We assumed that a slide labeled as MSI should contain at least a certain number (top K) of tiles having the highest MSI probabilities and vice versa. The same Densenet-IBN network was used for MIL training. In each epoch during training, all the tiles in one slide would be ranked according to their predicted probabilities, and only the top K tiles with the highest probabilities would then be used to calculate the loss and update model weights (Figure 1a). We empirically selected K equal to 20 in the study. Most of the training parameters and image augmentation methods were the same as fully supervised training, except that re-sampling of MSI samples was performed on the slide level other than tile level.

Generating MSI probability at slide level

For TCGA training and validation set, the tumor tiles provided by Kather were directly predicted by the trained MSI classifier without preforming tissue classification on WSI and a probability threshold of 0·5 was used to classify each tile into MSI or MSS. The percentage of tumor tiles classified as MSI in each slide was then calculated as the predicted probability of that slide.

In the external SYSUCC and PAIP test set, non-overlapping tiles with size of 224*224 pixels at 0·5 um/pixel were extracted from each WSI and classified into 9 classes with the trained tissue classifier first. The classification results were reviewed by a pathologist (YHL) to confirm suitability for further analysis. Non-overlapping tiles with the size of 512*512 pixels were then extracted from the identified tumor areas in the surgical samples. Since the sample size was relatively small in biopsy samples, 50% of overlapped tiles with the size of 512*512 pixels were extracted for all biopsy samples. The same approach of calculating the percentage of MSI tiles in each slide was used to generate the MSI probability at the slide level.

Comparisons with published MSI classifier

We compared our MSI classifier with a published MSI classifier using the fully supervised approach.[18] Specifically, model trained using four international cohorts with more than 6000 cases (identifier HLVDDREQHWQK, referred as Kather model in the following descriptions) was downloaded[33] and tested on our SYSUCC cohort.

Statistical analysis

The flowchart of our study was shown in Figure 1b. For assessing the cases characteristics, the mean (SD) was used for normally distributed continuous variables, and non-normally distributed data were presented as median (IQR) from the first (Q1) to the third (Q3) quartile. For categorical variables, percentages were reported. The receiver operating characteristic curves (ROCs) were drawn using GraphPad Prism v8·0· DeLong's test was used to evaluate statistically significant differences between the AUROC. A two-sided P value <0·05 was used to indicate statistical significance. We used the SPSS v24.0 for the analysis of AUROC, sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), F1-score and other statistical analysis.

Ethics statement

This study was approved by the Institutional Review Board (IRB) of SYSUCC (GZKJ2020-016). Informed consent from patients was waived due to the retrospective nature of this study.

Role of funding source

The funders had no role in the design of the study, in the collection, analyses, or interpretation of data, in the

writing of the manuscript, nor in the decision to publish the results.

Results

Patient characteristics

Baseline demographic information of the TCGA cohort, SYSUCC-surgical cohort, and SYSUCC-biopsy cohort are summarized in Table 1. The dominating ages of patients in the TCGA cohort, SYSUCC-surgical cohort, and SYSUCC-biopsy cohort were ≥50 years, with proportions of 88·4%, 68·5%, and 78·0%, respectively. The proportion of males in the TCGA cohort, SYSUCC-surgical cohort, and SYSUCC-biopsy cohort was 51·6%, 58·9%, and 63·0%, respectively. The dominating stages of all three groups were stage II and stage III. The dominating histological type was adenocarcinoma NOS with proportions of 88·0% (TCGA cohort), 77·5% (SYSUCC-surgical cohort), and 71.8% (SYSUCC-biopsy cohort). The proportion of dMMR were 17·5%, 24·8%, and 11·4% in the TCGA cohort, SYSUCC-surgical cohort, and SYSUCC-biopsy cohort, respectively. The majority of cases from the SYSUCC-surgical cohort and SYSUCC-biopsy cohort were located in left hemicolon or had no family history of CRC. The demographic information of the PAIP cohort was not available.

Predicting the MMR status of surgical specimens using deep learning models

The deep learning models to predict MSI-H were developed from three-fold cross-validation of the TCGA cohort. Two training strategies, including fully supervised (non-

Characteristics	TCGA *n* = 441	SYSUCC-surgical *n* = 355	SYSUCC-biopsy *n* = 341
Age, No. (%)			
<50	51(11.6)	112(31.5)	75(22.0)
≥50	389(88.4)	243(68.5)	266(78.0)
Missing	1	0	0
Gender, No. (%)			
Male	227(51.6)	209(58.9)	215(63.0)
Female	213(48.4)	146(41.1)	126(37.0)
Missing	1	0	0
AJCC stage (8th), No. (%)			
0	11(2.5)	5(1.4)	3(0·9)
1	69(15.7)	9(2.5)	54(15.8)
2	165(37.6)	252(71.0)	130(38.1)
3	129(29.4)	53(14.9)	106(31.1)
4	65(14.8)	36(10·1)	48(14.1)
Missing	2	0	0
Histology, No. (%)			
Adenocarcinoma NOS[a]	388(88.0)	275(77.5)	245(71.8)
Others[b]	53(12.0)	80(22.5)	96(28.2)
MMR status, No. (%)			
dMMR	77(17.5)	88(24.8)	39(11.4)
pMMR	364(82.5)	267(75.2)	302(88.6)
Location, No. (%)			
Left hemicolon	NA[c]	208(58.6)	253(74.2)
Right hemicolon	NA[c]	147(41.4)	88(25.8)
Family history, No. (%)			
Yes	NA[c]	96(27.0)	34(10·0)
No	NA[c]	259(73.0)	307(99.0)

***Table 1*: Baseline demographic information of the TCGA cohort, SYSUCC-surgical cohort and SYSUCC-biopsy cohort.**

TCGA = The Cancer Genome Atlas, SYSUCC = Sun Yet-sun University Cancer Center, pMMR = proficient mismatch repair, dMMR = deficient mismatch repair.

[a] Adenocarcinoma not otherwise specified (NOS), is a malignant epithelial salivary gland tumor with glandular or ductal adenocarcinomatous differentiation but without other specific histologic features, allowing for a more definitive classification and that characterizes the other defined types of salivary carcinoma.

[b] Histology subtypes other than adenocarcinoma NOS, such as mucinous adenocarcinoma and signet-ring cell carcinoma.

[c] Location and Family history are not included in TCGA cohort.

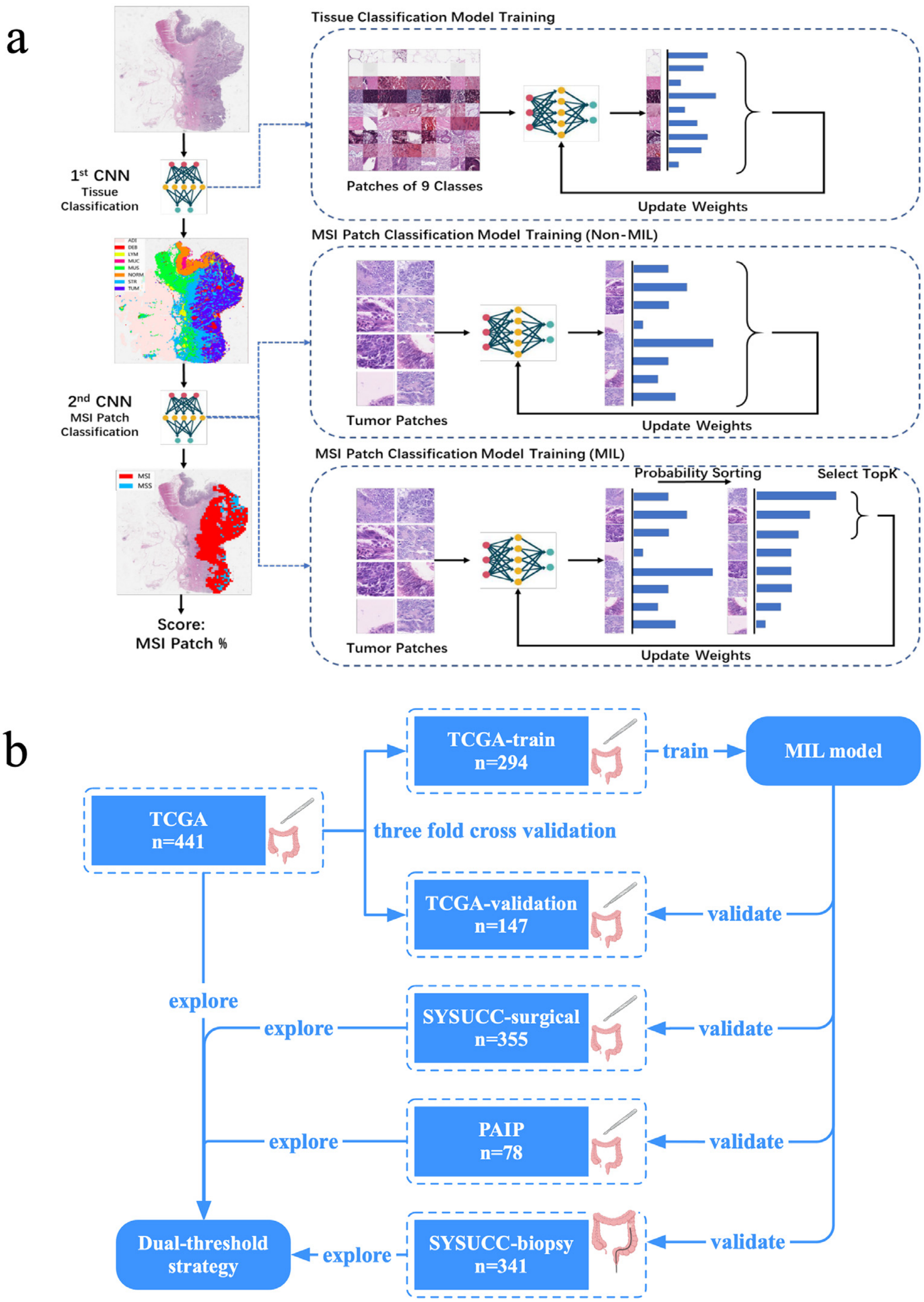

Figure 1. The workflow of deep learning model development (a) and the flowchart of our study (b).

(a) A tissue classification model was first developed to recognize nine different tissue structures from H&E stained WSIs. The MSI/MSS classification model was trained from patches extracted from identified tumor areas. Two training strategies, including fully supervised training and MIL training, were investigated. The percentage of classified MSI patches among all the tumor patches was calculated as the final score.

(b) Flowchart of all experiments. The MIL model was trained in the TCGA-training cohort and validated in the TCGA-validation, SYSUCC-surgical, SYSUCC-biopsy cohorts. The dual-threshold strategy was explored in the surgical cohort and validated in the SYSUCC-biopsy cohort.

WSI: Whole slide image; MIL = multiple instance learning, MSI = Microsatellite instability, CNN = Convolutional Neural Network, TCGA = The Cancer Genome Atlas, SYSUCC = Sun Yet-sun University Cancer Center.

MIL) and weakly supervised (MIL), were adopted. The overall average AUROC was defined as the average AUROC from the three folds. As shown in Figure 2 and Supplementary Table S3, the MIL model had similar average AUROC (0·9726±0·0156) as the non-MIL model (0·9678±0·0172) of the TCGA-train cohort. Slight degradations in the performance of both models were found in the TCGA-validation cohort. As shown in Figure 2 and Supplementary Table S3, the MIL model had an average AUROC of 0·8888±0·0357, and the non-MIL model had an average AUROC of 0·8830±0·0540.

As the tissue samples provided in the TCGA cohort were mainly surgical specimens, we first performed an external validation using two independent surgical specimen cohorts (SYSUCC-surgical cohort and PAIP cohort). Although the patient ethnicity, sample processing protocols and image scanning machines were different between the TCGA, PAIP and SYSUCC cohorts, the average AUROC of the MIL model of the SYSUCC-surgical cohort (0·8457±0·0233) and PAIP cohort (0·8806±0·0232) was robust, compared to the TCGA-validation cohort (Figure 2 and Supplementary Table S3). The non-MIL method could also predict MSI-H status but had the worse average AUROC in the SYSUCC-surgical cohort (0·8133±0·0479) and the PAIP cohort (0·8412±0·0182). It should be noted that the MIL model performed better than the non-MIL model in the fold 1 (0·8191 vs. 0·7611, P = 0·004, DeLong's test) and fold 3 (0·8628 vs. 0·8235, P = 0·026, DeLong's test) of the SYSUCC-surgical cohort.

The Kather model, established using the largest CRC patient cohort up-to-date from a combination of four cohorts with more than 6000 cases, was obtained from the Echle's study to compare the SYSUCC-surgical cohort with the PAIP cohort.[18] The Kather model was trained using a supervised training strategy which was similar to our non-MIL model. Even with the advantage of a much larger training dataset size, the Kather model only yielded a similar AUROC of 0·8517 (95% CI: 0·7993−0·9040) in the SYSUCC-surgical cohort and 0·9117 (95% CI: 0·8327−0·9907) in the PAIP cohort compared to the MIL and non-MIL models (Figure 2 and Supplementary Table S3).

Predicting the MMR status of biopsy specimens using deep learning models

We next validated the models on another independent biopsy specimen cohort (SYSUCC-biopsy). Compared with the surgical specimens, the tumor area identified in the biopsy specimens was much smaller which posed difficult challenges. As shown in Figure 2 and Supplementary Table S4, the AUROC of MIL model was significantly better than the non-MIL model (P = 0·686 in fold 1, P < 0·001 in fold 2, P = 0·009 in fold 3, DeLong's test) and Kather model (P = 0·035 in fold 1, P < 0·001 in fold 2, P = 0·002 in fold 3, DeLong's test). The average AUROC of the MIL model could improve from 0·7679±0·0342 to 0·7849±0·0312 if the biopsies with minimal tumor area detected (less than 100 patches) were excluded (Supplementary Figure S3). The baseline characteristics of low tumor tile biopsies were shown in Supplementary Table S5.

Subgroup analysis of the MIL model

Since the MIL model performed best in both the SYSUCC-surgical and SYSUCC-biopsy cohorts, further subgroup analysis was performed to validate the robustness of the MIL model in regard to the subgroup's heterogeneity. In the SYSUCC-surgical cohort, variation of AUROC was significant only in fold 3 between adenocarcinoma, not otherwise specified (NOS) and other subtypes (AUROC=0·8899 vs 0·7418, P = 0·0329, DeLong's test). Less variations were observed in other fold or between other subgroup such as age, genders, locations, AJCC stages, and family history (Supplementary Figure S5-7). The AUROC variations between different subgroup were also found not significant in SYSUCC-biopsy cohort except that the AUROC of the adenocarcinoma NOS in fold 1 was better when compared with the other subtypes (AUROC=0·7832 vs 0·5428, P = 0·0021, DeLong's test, Supplementary Figure S8-10). In general, there existed less variations in model performance between various subgroups, and the overall performance was stable except that the model performance reduced for histology subtypes other than adenocarcinoma NOS. Additionally, we explored the variations between lynch syndrome and sporadic dMMR. Thirty-one and twenty patients in the SYSUCC-surgical cohort were confirmed as lynch syndrome and sporadic dMMR through separate genetic test. A significant difference in predicted scores from the MIL model could be observed between these lynch syndrome cases and the sporadic dMMR cases in fold 3 (P = 0·01), implying the possible differentiation of lynch syndrome from dMMR using MIL model (Supplementary Figure S11).

Dual-threshold strategy for patient triage

Applying quantitative rule-in and rule-out thresholds for predicted scores of MIL model could separate the patients into confirmed dMMR, uncertain and confirmed pMMR groups, which could reduce unnecessary IHC or genetic dMMR tests. The main evaluation indicators of our dual-threshold strategy were sensitivity, specificity, PPV, NPV, F1-score and IHC rate. We aimed to keep a high sensitivity and specificity, and a low IHC rate as far as possible. The predicted scores from three-fold models were first averaged. Both thresholds were identified from each cohort based on the sensitivity and specificity, which were then evaluated by the corresponding PPV, NPV, F1-score and IHC rate. The rule-out threshold (lower threshold) was used to recognize

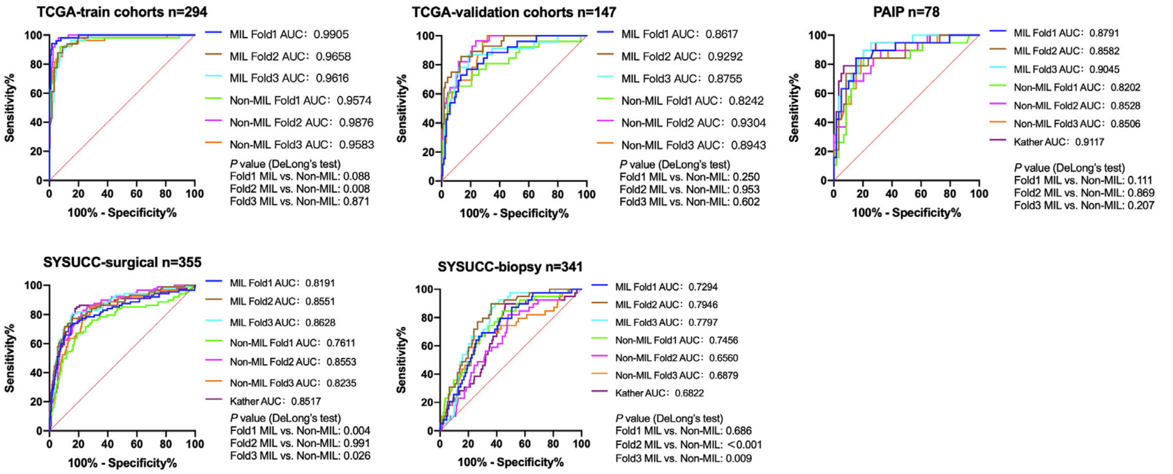

Figure 2. The ROC curves, the AUROC, and the corresponding *P* value of MIL model and other models on the TCGA-train cohort, TCGA-validation cohort, SYSUCC-surgical cohort, SYSUCC-biopsy cohort and PAIP cohort.

ROC curves and respective AUROCs of MIL model, Non-MIL model, Kather model are presented for MMR detection on the TCGA-train cohort, TCGA-validation cohort, PAIP cohort, SYSUCC-surgical cohort and SYSUCC-biopsy cohort. DeLong's test was used to evaluate statistically significant differences between the AUROC.

AUROC=area under the ROC curve, ROC=receiver operating characteristic, TCGA = The Cancer Genome Atlas, SYSUCC = Sun Yet-sun University Cancer Center, MMR = mismatch repair.

pMMR patients who would not benefit from IHC or genetic tests and their sensitivity of dMMR detection was kept at higher than 90%. On the other hand, the rule-in threshold (upper threshold) was designed to identify as many dMMR patients who unnecessarily received IHC or genetic tests, while keeping the specificity of dMMR detection higher than 95%. No further testing was required for the patient if the predicted score from the MIL model was above the rule-in threshold or below the rule-out threshold. If the predicted score was between the dual thresholds, the patient would be suggested to receive further IHC or genetic tests to confirm the MMR status. Given the optimized thresholds identified above, the sensitivity, specificity, PPV, NPV, F1-score, and IHC rate were 90·91%, 95·05%, 79·55%, 98·02%, 84·85% and 4·54% in the TCGA cohort, 90·91%, 95·13%, 86·02%, 96·95%, 88·40% and 43·94% in the SYSYCC-surgical cohort, and 94·74%, 96·61%, 90·00%, 98·28%, 92·31% and 47·44% in the PAIP cohort (Table 2). When the dual-threshold strategy was applied to the SYSUCC-biopsy cohort, dMMR detection using the MIL model achieved a sensitivity of 92·31%, specificity of 95·36%, 72·00% in PPV, 98·97% in NPV, and 80·90% in F1-score, with 48·97% of the patients requiring further testing (Table 2).

Discussion

In this study, we demonstrated that the deep learning models developed using domain adaptive and weakly supervised training techniques could predict MMR status directly from H&E stained tissue specimens. An optimized dual-threshold strategy was established, which was able to facilitate the potential adoption of artificial intelligence (AI)-digital pathology-based MMR status triage with high sensitivity and specificity.

Great success in predicting MMR status from digital pathology images has been reported in several

	TCGA *n* = 441	SYSUCC-surgical *n* = 355	SYSUCC-biopsy *n* = 341	PAIP *n* = 78
Sensitivity	0·9091	0·9091	0·9231	0·9474
Specificity	0·9505	0·9513	0·9536	0·9661
PPV	0·7955	0·8602	0·7200	0·9000
NPV	0·9802	0·9695	0·9897	0·9828
F1-score	0·8485	0·8840	0·8090	0·9231
IHC rates	0·0454	0·4394	0·4897	0·4744

***Table 2*: The evaluation index of dual-threshold strategy on TCGA cohort, SYSUCC-surgical cohort, SYSUCC-biopsy cohort, and PAIP cohort.**

PPV = negative predictive value, NPV = negative predictive value, IHC = immunohistochemistry.

studies.[18,20,21] One widely accepted approach was fully supervised, which assumed the MMR status was uniform in the one WSI, and the same slide-level label was assigned to every tile inside for training.[31] Such assumption ignored the heterogeneity of dMMR and pMMR pattern distributions in the WSI, which could affect the deep learning model training. A weakly supervised training approach, especially MIL, has been shown to be highly effective in other AI-digital pathology tasks such as cancer detection,[26] prognosis,[29] or tumor subtyping[30] with only slide-level labels available. Two recent studies have successfully utilized MIL for MSI detection in colorectal cancer and demonstrated its superiority over fully supervised-based approaches.[23,34] Schirris et al.[34] embedded each image patch into a feature vector using contrastive self-supervised learning approaches and aggregated all vectors to produce slide-level prediction with a modified attention-based MIL. Another MIL approach by Bilal et al.[23] was to train an instance-level classifier according to top K instances that are most likely to have the same label as the WSI. Our method was similar to Bilal's method without adopting a training strategy such as iterative draw, but we incorporated a domain adaption layer to strengthen the model generalization ability. Similar to the findings reported in these two studies, our MIL model was superior to the non-MIL model in both the TCGA validation and SYSUCC-surgical cohorts. Moreover, the AUROC of the MIL model in these two cohorts was almost the same (0·88 vs. 0·85), showing an even better model generalizability. It should also be noted that our MIL model achieved similar performance as the Kather model in the SYSUCC-surgical cohort. The size of our training set was only one twentieth the size used in Kather's study (n = 294 vs. n = 6404).[18] With the growing training size, our model would have the potential to further improve. The MIL technique could also be improved with the recent advances of attention based[35] or knowledge-distill based[36] methods.

Identifying the MSI status in biopsy specimens has more significant clinical impact but is also more challenging. Almost every CRC patient would have the available biopsy specimen from endoscopy or fine-needle aspiration to confirm the diagnosis. Detection of MSI status from biopsy samples allows earlier triage of patients, reduces unnecessary IHC or genetic testing in some patients, and allows earlier initiation of the corresponding treatment. However, biopsy specimens could undergo blunt extrusion and tearing during sample extraction, leading to architectural distortion and variable alteration. Additionally, previous study pointed that analyses involving single-site biopsy sampling might result in underestimation of the degree of spatial heterogeneity.[37] Echle et al. reported an AUROC of 0·78 by applying the Kather model directly on an independent biopsy cohort.[18] Although the same model performed poorly on our biopsy cohort (AUROC=0·68), our MIL model still achieved an acceptable performance (AUROC=0·77) on biopsy samples demonstrating the advantage of MIL method over fully supervised training techniques. If the biopsy specimens with less than 100 tumor tiles were excluded, the AUROC could be improved to 0·7849. Although the difference is not statistically significant, which may due to the small sample size (low tumor tile biopsies, n = 29), we can still find some differentiated trends. This result indicated the importance of tumor area adequacy in the biopsy sample for MMR detection, and a tissue quality assurance step might be mandatory in the workflow.

To assess whether the performance of the MIL model is robust across heterogeneous subgroups, we compared AUROC across subgroups in the SYSUCC-surgical and SYSUCC-biopsy cohorts. The subgroup analysis showed less variations across clinicopathological features such as age, anatomical location, AJCC stage and family history of cancer. Our result was similar to Echle's study and demonstrated the robustness of the MIL model.[18] However, we also found significant variation between adenocarcinoma NOS and other histology subtypes in fold 3 in the SYSUCC-surgical cohort and in fold1 in the SYSUCC-biopsy cohort. It is probably because the majority of cases in the training cohort were adenocarcinoma. Adding more cases with other histology subtypes into the training cohort in the future could effectively improve the model's performance in this subgroup. Additionally, we found that the predicted score of Lynch syndrome was higher than those of sporadic dMMR CRCs. It would be interesting to develop a specific model for Lynch syndrome detection from H&E samples in future studies.

Identifying appropriate thresholds for the predicted scores from deep learning models is not trivial, especially for the actual clinical application. A single threshold strategy was commonly applied for surgical specimens to identify sensitivity and specificity by selecting optimal threshold from statistical approaches such as Youden's J statistic and the weighted Youden index[20] or studying various applicable threshold values.[38] However, the current single threshold strategy failed to achieve high sensitivity and high specificity for dMMR detection at the same time and was not validated on biopsy samples. To address this issue, Kacew et al did a theoretical comparison between eight testing strategies, with different thresholds selected favoring either high-sensitivity or high-sensitivity for AI models.[38] The high-sensitivity AI model (sensitivity=98%, specificity = 79%) followed by high-specificity IHC or genetic testing panel was found to be the best strategy to lower the total testing and first-line drug therapy cost. It is similar to our dual-threshold strategy, although these performances were generated from surgical specimens only, while our dual threshold strategy could work prominently on biopsy specimens.

The workflow of dual-threshold strategy is shown in Figure 3. For the confirmed dMMR group identified

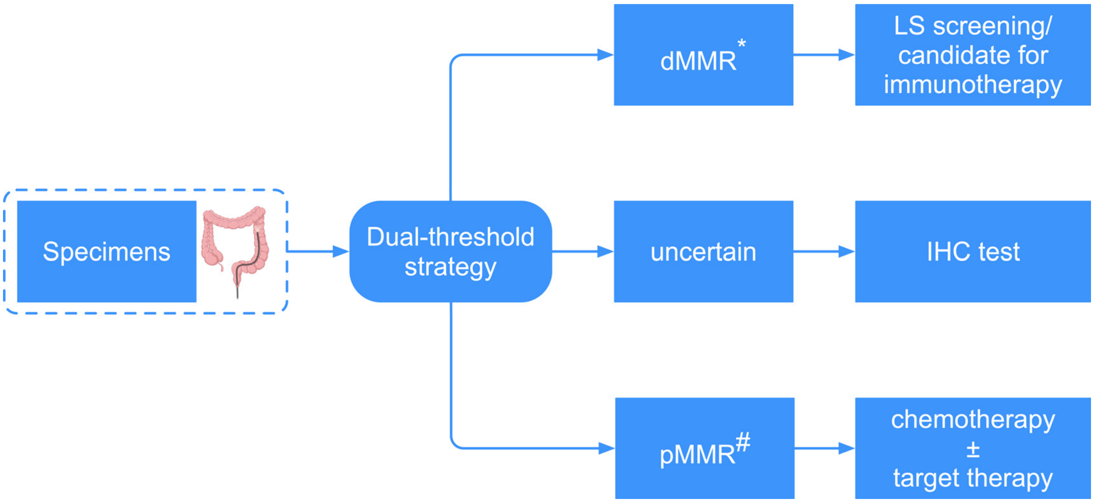
Specimens
Dual-threshold strategy
dMMR*
uncertain
pMMR#
LS screening/ candidate for immunotherapy
IHC test
chemotherapy ± target therapy

Figure 3. Workflow for MMR screening with dual-threshold strategy.

The new screening strategy, with the help of our dual-threshold approach, is shown in the figure. The biopsy specimens with H&E stain can be transferred into WSIs through a scanning electron microscope. Subsequently, the WSIs were sent to the MIL model, and a predicted result was given. Only the suspicious group would require further IHC tests to confirm dMMR status.

* Set the up-threshold higher to improve specificity and minimize false dMMR.

Patient suspected to be dMMR clinically (family history, multiple primary tumors, etc) would receive salvage IHC test.

H&E = haematoxylin and eosin, WSI = whole-slide image, IHC = immunohistochemistry, pMMR = proficient mismatch repair, dMMR = deficient mismatch repair, LS = lynch syndrome.

with predicted scores above the upper threshold, the patients could be referred to germline testing for LS screening or to receive tumor sequencing from identifying genetic mutation and tumor mutation burden (TMB) status for the choice of immunotherapy. Only the uncertain group would require further IHC/PCR testing to confirm the dMMR status. We evaluated the dual threshold strategy on TCGA cohort, SYSUCC-surgical cohort, SYSUCC-biopsy cohort, and PAIP cohort, and we succeeded in keeping the sensitivity and specificity higher than 90% and 95% while less than half of the patients needed to receive further IHC/PCR based dMMR testing. The two thresholds could be adjusted in the larger cohort study to further decrease the percentage of patients requiring additional testing to further lowering of the average cost.

One limitation of our study is that our models were trained from surgical specimens and tested on biopsy specimens despite the heterogeneity between them. We selected CRC surgical specimens from the TCGA cohort as the training samples for the direct comparisons with published models using the same dataset. Although we have shown that the deep learning model trained from surgical specimens using MIL could be generalized to biopsy specimens, the performance of the models on biopsy specimens was still inferior to the performance on surgical specimens. Similar findings were reported by Echle A et al.[18]. They claimed that the performance on biopsy specimens could be improved by training directly from biopsy specimens using the non-MIL method. Since not enough biopsy specimens were recruited in this study, to validate such claim, we could only conclude from our findings that models trained from the MIL method might be more robust to sampling variation and provide better clinical utility on biopsy specimens than the non-MIL method when only surgical specimens were available for training, as proposed in the majority studies. Whether training from biopsy specimens using MIL could improve the performance needs further investigation. Secondly, the sizes of our training and testing cohorts were relatively small and were from retrospective analyses which limited the significance of the results. The sample size issue may matter more when the MIL technique was adopted as each slide was treated as one sample during training, while each patch was treated as one sample during the fully supervised training, but thousands of patches could be generated from one slide to increase sample size significantly. Hence the MIL training strategy could benefit significantly from future studies with more samples collected. More importantly, not only the model performance could be improved, but more robust thresholds could be obtained from larger sample size. The potential clinical applicability of the dual thresholds identified in this study might be limited by the sample size, but our study successfully established a useful CRC patient triage strategy to detect tumor MMR status directly from H&E stained tissue slide, especially from biopsy specimens. Such a workflow could be more useful in a primary care hospital and developing countries with little access to MMR IHC or PCR. However, it should be noted that our test materials were not from these test sites while factors such as fixation, quality of glass and H&E staining could be different in such areas, and we would like to validate our model in these environments in the future.

In summary, we developed a deep learning-based system to predict CRC MMR status from H&E stained pathology images which is robust between different centers and performs well on both surgical and biopsy specimens. A dual-threshold triage strategy was demonstrated to supplement the current screening workflow by excluding over half of the CRC patients for further IHC/PCR-based MMR testing while maintaining high sensitivity and specificity of dMMR detection.

Contributors

WJ, WJM, SYX and YHL contributed equally to this manuscript and are listed as co-first authors; SYX, JBK, HSL performed the majority of the machine learning development and validation with input from WJ, WJM, YHL and WRL; SYX and HH wrote the technical infrastructure needed for machine learning; WJ, WJM and YHL collected and performed quality control for the data; HZZ, LL and PRD verified the underlying data. JBL, MYC, ZZP, HZZ, LL and PRD provided critical feedback and edited the manuscript. WJ, WJM, SYX and YHL wrote the manuscript with assistance and feedback from all the other co-authors. All authors read and approved the final version of the manuscript. WJ and PRD accessed and verified the data and were responsible for the decision to submit the manuscript.

Data sharing statement

Due to the privacy of patients, the related data cannot be available for public access but can be obtained from Pei-Rong Ding (dingpr@sysucc.org.cn) upon reasonable request. We have made our codes open access at https://github.com/biototem/MSI_classifier.

Declaration of interests

SYX and HH are cofounders of Bio-totem. JBK and HSL are full-time employees at Bio-totem. No disclosures were reported by the other authors.

Acknowledgements

We are grateful to Yiducloud (Beijing) technology Ltd. for providing data extraction and processing platform for this research. De-identified pathology images and annotations used in this research were prepared and provided by the Seoul National University Hospital by a grant of the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea (grant number: HI18C0316). The study was funded by the National Natural Science Foundation of China (82073159, 81871971 and 81700576), the Natural Science Foundation of Guangdong Province (No. 2021A1515011792 and No. 2022A1515012403) and Medical Scientific Research Foundation of Guangdong Province of China (No. A2020392).

Supplementary materials

Supplementary material associated with this article can be found in the online version at doi:10.1016/j.ebiom.2022.104120.

References

1 Sung H, Ferlay J, Siegel RL. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin*. 2021;71(3):209–249.

2 Siegel RL, Miller KD, Fuchs HE, Jemal A. Cancer statistics, 2021. *CA Cancer J Clin*. 2021;71(1):7–33.

3 Serebriiskii IG, Connelly C, Frampton G, et al. Comprehensive characterization of RAS mutations in colon and rectal cancers in old and young patients. *Nat Commun*. 2019;10(1):3722.

4 Latham A, Srinivasan P, Kemel Y, et al. Microsatellite instability is associated with the presence of lynch syndrome pan-cancer. *J Clin Oncol*. 2019;37(4):286–295.

5 Petrelli F, Ghidini M, Ghidini A, Tomasello G. Outcomes following immune checkpoint inhibitor treatment of patients with microsatellite instability-high cancers: a systematic review and meta-analysis. *JAMA Oncol*. 2020;6(7):1068–1071.

6 Sargent DJ, Marsoni S, Monges G, et al. Defective mismatch repair as a predictive marker for lack of efficacy of fluorouracil-based adjuvant therapy in colon cancer. *J Clin Oncol*. 2010;28(20):3219–3226.

7 Sinicrope FA. The role of microsatellite instability testing in management of colorectal cancer. *Clin Adv Hematol Oncol*. 2016;14 (7):476–479.

8 Luchini C, Bibeau F, Ligtenberg MJL, et al. ESMO recommendations on microsatellite instability testing for immunotherapy in cancer, and its relationship with PD-1/PD-L1 expression and tumour mutational burden: a systematic review-based approach. *Ann Oncol*. 2019;30(8):1232–1243.

9 NCCN Guidelines Version 1.2022 Genetic/Familial High-Risk Assessment: Colorectal. Available at: https://www.nccn.org/professionals/physician_gls/pdf/genetics_colon.pdf. Accessed 8 June 2022.

10 Shia J. The diversity of tumours with microsatellite instability: molecular mechanisms and impact upon microsatellite instability testing and mismatch repair protein immunohistochemistry. *Histopathology*. 2021;78(4):485–497.

11 Vilar E, Gruber SB. Microsatellite instability in colorectal cancer-the stable evidence. *Nat Rev Clin Oncol*. 2010;7(3):153–162.

12 Halvarsson B, Anderson H, Domanska K, Lindmark G, Nilbert M. Clinicopathologic factors identify sporadic mismatch repair-defective colon cancers. *Am J Clin Pathol*. 2008;129(2):238–244.

13 Brazowski E, Rozen P, Pel S, Samuel Z, Solar I, Rosner G. Can a gastrointestinal pathologist identify microsatellite instability in colorectal cancer with reproducibility and a high degree of specificity? *Fam Cancer*. 2012;11(2):249–257.

14 Jiang Y, Yang M, Wang S, Li X, Sun Y. Emerging role of deep learning-based artificial intelligence in tumor pathology. *Cancer Commun (Lond)*. 2020;40(4):154–166.

15 Acs B, Hartman J. Next generation pathology: artificial intelligence enhances histopathology practice. *J Pathol*. 2020;250(1):7–8.

16 Mori Y, Bretthauer M, Kalager M. Hopes and hypes for artificial intelligence in colorectal cancer screening. *Gastroenterology*. 2021;161(3):774–777.

17 Chen ZH, Lin L, Wu CF, Li CF, Xu RH, Sun Y. Artificial intelligence for assisting cancer diagnosis and treatment in the era of precision medicine. *Cancer Commun (Lond)*. 2021;41(11):1100–1115.

18 Echle A, Grabsch HI, Quirke P, et al. Clinical-grade detection of microsatellite instability in colorectal tumors by deep learning. *Gastroenterology*. 2020;159(4):1406.

19 Thakur N, Yoon H, Chong Y. Current trends of artificial intelligence for colorectal cancer pathology image analysis: a systematic review. *Cancers (Basel)*. 2020;12(7):1884.

20 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(1):132–141.

21 Bilal M, Raza SEA, Azam A, et al. Development and validation of a weakly supervised deep learning framework to predict the status of molecular pathways and key mutations in colorectal cancer from routine histology images: a retrospective study. *Lancet Digit Health*. 2021;3(12):e763–e772.

22 Echle A, Ghaffari Laleh N, Quirke P, et al. Artificial intelligence for detection of microsatellite instability in colorectal cancer-a multicentric analysis of a pre-screening tool for clinical application. *ESMO Open*. 2022;7(2):100400.

23 Bilal M, Raza SEA, Azam A, et al. Development and validation of a weakly supervised deep learning framework to predict the status of molecular pathways and key mutations in colorectal cancer from routine histology images: a retrospective study. *Lancet Digit Health*. 2021;3(12):e763–ee72.

24 Cao R, Yang F, Ma SC, et al. Development and interpretation of a pathomics-based model for the prediction of microsatellite instability in colorectal cancer. *Theranostics*. 2020;10(24):11080–11091.

25 Echle A, Laleh NG, Schrammen PL, et al. Deep learning for the detection of microsatellite instability from histology images in colorectal cancer: a systematic literature review. *ImmunoInformatics*. 2021;3-4:100008.

26 Kather JN. Image tiles of TCGA-CRC-DX histological whole slide images, non-normalized, tumor only (v0.1) [Data set]. Zenodo. 2020. https://doi.org/10.5281/zenodo.3784345.

27 https://paip2020.grand-challenge.org/Dataset/. Accessed 18 June 2022.

28 Kather JN, Halama N, Marx A. 100,000 histological images of human colorectal cancer and healthy tissue (v0.1) [Data set]. *Zenodo*. 2018. https://doi.org/10.5281/zenodo.1214456.

29 Xu J, Cai C, Zhou Y, et al. Multi-tissue partitioning for whole slide images of colorectal cancer histopathology images with deeptissue net. *European Congress on Digital Pathology*. Cham; Springer; 2019:100–108.

30 Pan X, Luo P, Shi J, Tang X. Two at once: Enhancing learning and generalization capacities via ibn-net. In: *Proceedings of the European Conference on Computer Vision (ECCV)*. 2018464–479.

31 Kather JN, Pearson AT, Halama N, et al. Deep learning can predict microsatellite instability directly from histology in gastrointestinal cancer. *Nat Med*. 2019;25(7):1054–1056.

32 Campanella G, Hanna MG, Geneslaw L, et al. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nat Med*. 2019;25(8):1301–1309.

33 Kather JN. Deep learning models to detect microsatellite instability in colorectal cancer from histological images (v0.1) [Data set]. Zenodo. 2020. https://doi.org/10.5281/zenodo.3627523.

34 Schirris Y, Gavves E, Nederlof I, Horlings HM, Teuwen J. DeepSMILE Contrastive self-supervised pre-training benefits MSI and HRD classification directly from H&E whole-slide images in colorectal and breast cancer. *Med Image Anal*. 2022;79:102464.

35 Lu MY, Williamson DFK, Chen TY, Chen RJ, Barbieri M, Mahmood F. Data-efficient and weakly supervised computational pathology on whole-slide images. *Nat Biomed Eng*. 2021;5(6):555–570.

36 Ke J, Shen Y, Wright JD, Jing N, Liang X, Shen D. Identifying patch-level MSI from histological images of colorectal cancer by a knowledge distillation model. *2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE; 2020:1043–1046.

37 Dagogo-Jack I, Shaw AT. Tumour heterogeneity and resistance to cancer therapies. *Nat Rev Clin Oncol*. 2018;15(2):81–94.

38 Lee SH, Song IH, Jang HJ. Feasibility of deep learning-based fully automated classification of microsatellite instability in tissue slides of colorectal cancer. *Int J Cancer*. 2021;149(3):728–740.