Predicting microsatellite instability and key biomarkers in colorectal cancer from H&E-stained images: Achieving SOTA with Less Data using Swin Transformer

Author: Bangwei Guo¹, Jitendra Jonnagaddala², Hong Zhang³*, Xu Steven Xu⁴*

¹Department of Applied Statistics, School of Data Science, University of Science and Technology of China;  
²School of Population Health, UNSW Sydney, Kensington, NSW, Australia  
³Department of Statistics and Finance, School of Management, University of Science and Technology of China;  
⁴Data Science/Translational Research, Genmab Inc., Princeton, New Jersey, USA;
Abstract

Artificial intelligence (AI) models have been developed for predicting clinically relevant biomarkers, including microsatellite instability (MSI), for colorectal cancers (CRC). However, the current deep-learning networks are data-hungry and require large training datasets, which are often lacking in the medical domain. In this study, based on the latest Hierarchical Vision Transformer using Shifted Windows (Swin-T), we developed an efficient workflow for biomarkers in CRC (MSI, hypermutation, chromosomal instability, CpG island methylator phenotype, \textit{BRAF}, and \textit{TP53} mutation) that only required relatively small datasets, but achieved the state-of-the-art (SOTA) predictive performance. Our Swin-T workflow not only substantially outperformed published models in an intra-study cross-validation experiment using TCGA-CRC-DX dataset (\(N = 462\)), but also showed excellent generalizability in cross-study external validation and delivered a SOTA AUROC of 0.90 for MSI using the MCO dataset for training (\(N = 1065\)) and the same TCGA-CRC-DX for testing. Similar performance (AUROC = 0.91) was achieved by Elche and colleagues using ~8000 training samples (ResNet18) on the same testing dataset. Swin-T was extremely efficient using small training datasets and exhibits robust predictive performance with only 200 – 500 training samples. These data indicate that Swin-T may be 5–10 times more efficient than the current state-of-the-art algorithms for MSI based on ResNet18 and ShuffleNet. Furthermore, the Swin-T models showed promise as pre-screening tests for MSI status and \textit{BRAF} mutation status, which could exclude and reduce the samples before the subsequent standard testing in a cascading diagnostic workflow to allow turnaround time reduction and cost saving.
Introduction

Artificial-intelligence (AI) and Deep-learning models using hematoxylin and eosin (H&E) stained histology slides have been developed to predict clinically relevant molecular biomarkers for colorectal cancer (CRC) such as microsatellite instability (MSI)\textsuperscript{1-3} genetic mutations\textsuperscript{4-6}, molecular subtypes\textsuperscript{6,7}, etc. Particularly, prediction of MSI is of great clinical utility since MSI is the only approved biomarker to select patients for immune checkpoint inhibitors\textsuperscript{8}. The US Food and Drug Administration (FDA) granted accelerated approval to anti-PD1 antibodies (e.g., Pembrolizumab and Nivolumab, etc.) for treatment of MSI-High or mismatch repair deficient (dMMR) cancers, including CRC (the FDA’s first tissue/site-agnostic approval)\textsuperscript{9,10}.

Kather et al. developed the first fully automated Deep-Learning (DL) model for predicting MSI/dMMR status in CRC in 2019\textsuperscript{1}. Since then, multiple models have been published\textsuperscript{1-4,6,11-16}. Recently, Vision Transformer (ViT)\textsuperscript{17} has emerged as a brand-new modeling framework in the field of computer vision and show great potential to replace Convolutional Neutral Network (CNN)\textsuperscript{14,17}, which has been the backbone for the vast majority of DL models in digital pathology (including MSI models). The earlier versions of ViT still require very large datasets to achieve comparable performance as CNN models\textsuperscript{17,18}.

In the medical imaging domain, datasets are usually limited and often accompanied by weak (i.e., slide-level) labels\textsuperscript{19}. Also, attention-based DL models such as ViT are always complex and data-hungry in nature\textsuperscript{19,20}. These challenges pose a huge barrier for development and application of DL models in digital pathology. For example, the current SOTA for predicting MSI status was achieved using extremely large, pooled datasets from different studies (N > 8,000 WSIs)\textsuperscript{11}. Therefore, how to achieve clinical-grade prediction of MSI status and other key biomarkers for CRC with limited data remains an active research question. The latest Hierarchical Vision
Transformer using Shifted Windows (Swin Transformer; Swin-T thereafter) reduces the computational complexity and can flexibly process pictures with different scales\textsuperscript{18}. Therefore, Swin Transformer may have the potential to circumvent the limitations of small datasets in medical image studies.

In this study, we aimed to develop an efficient workflow using Swin-T that can use relatively small training datasets, but achieve the best, state-of-the-art (SOTA) predictive performance for MSI status and other key biomarkers in CRC (BRAF mutation, TP53 mutation, CpG island methylator phenotype (CIMP), hypermutation, and chromosomal instability) using hematoxylin and eosin (H&E) stained images of colorectal tumors.

**Method**

The workflow for processing the Whole Slide Images (WSIs) and modeling the data is illustrated in Figure 1. In this study, we developed a novel Swin-T based deep learning pipeline for predicting key biomarkers for CRC patients including MSI-status. This pipeline included 2 Swin-T models: a tissue classifier to detect tumor tissues and a biomarker classifier to predict binary biomarker status.

**Imaging and clinical data**

Two international CRC datasets are analyzed in this study. The MCO (Molecular and Cellular Oncology) dataset is a collection of any stage patients who underwent curative resection for colorectal cancer between 1994 and 2010 in New South Wales, Australia. The public TCGA dataset (‘The Cancer Genome Atlas’, publicly available at https://portal.gdc.cancer.gov/, USA) is
a multi-centric collection of tissue specimens, which includes tumors of all stages in the TCGA-COAD and TCGA-READ datasets.

Anonymized hematoxylin and eosin (H&E) stained whole-slide images are collected from two datasets with matched genomic data. For the MCO dataset, a small number of patients (n = 73) are excluded because of absence of tumor tiles and lack of molecular information. After excluding, 1065 WSIs from 1065 patients from the MCO dataset were used in the present study. The ground truth labels of the MCO dataset are available for MSI status (157 microsatellite instable and 908 microsatellite stable), *BRAF* mutation (117 mutants and 909 wild-type), and CIMP (153 CIMP-high and 211 CIMP-low). The TCGA-CRC-DX dataset, which has been used in multiple previously published studies for MSI prediction4,11,12,14, includes 502 whole-slide images of primary colorectal tumors from 499 patients. A summary of the ground truth labels in the TCGA-CRC-DX dataset for different biomarkers (hypermutation, microsatellite instability, chromosomal instability, CIMP, *BRAF*, and *TP53* mutation) is available in Bilal et al4.

**Data preprocessing**

Scanned WSIs are downloaded in SVS format. All WSIs are tessellated into small image tiles of 512×512 pixels at a resolution 0.5 μm. No manual annotations of tumor tissue are used. The image tiles were color normalized using Macenko’s method21 to reduce the color bias and improve classifier performance, and were then resized to 224×224 pixels to serve as the input of the network. Image tiles containing background or blurry ones were automatically removed from the dataset during this process using the detected edge quantity (canny edge detection in Python’s OpenCV package) (https://github.com/KatherLab/preProcessing).

A Swin-T tissue classifier is trained to detect and select tiles with tumor tissue using a publicly available dataset, NCT-CRC-HE-100K, which consists of nine types of CRC tissue
images collected from Kather et al. Up to 500 tumor tiles are randomly selected per patient, which are used for all subsequent steps. And all selected tiles inherit the label of the corresponding patient so all models are trained only using slide-level labels.

**Training strategy for deep learning models**

A stepwise strategy was adopted during model development.

**Pre-training a Swin-T tissue classifier**

First, a Swin Transformer model was pre-trained to develop a multi-class tissue classifier. The tissue classifier was trained and tested using two publicly available, pathologist annotated datasets (NCT-CRC-HE-100K and CRC-VAL-HE-7K, respectively) from Kather and colleagues. These 2 datasets consist of CRC image tiles of nine tissue types: adipose tissue (ADI), background (BACK), debris (DEB), lymphocytes (LYM), mucus (MUC), smooth muscle (MUS), normal colon mucosa (NORM), cancer-associated stroma (STR), and colorectal adenocarcinoma epithelium (TUM). The NCT-HE-100K dataset has 100,000 image tiles (TUM = 14317 and non-tumor = 85683). All the image patches were 224 x 224 pixels at 20X magnification. During pre-training process, the model was trained for 10 epochs using a fixed learning rate of 0.001. Adam optimizer and cross entropy loss function were used. After the pre-training was completed, the parameter weights of the backbone part of the pretrained model were saved. The Swin-T tissue classifier achieved a high overall accuracy (96.3%) and tumor detection accuracy (98%) of the tissue segmentation model on the unseen test set, CRC-VAL-HE-7K, which contains 7180 image tiles (TUM = 1223 and non-tumor = 6957) (Supplementary Figure 1).

**Fine-tuning Swin-T biomarker models**

Then, the pre-trained Swin Transformer model (tissue classifier) was fine-tuned for binary
classification for key CRC biomarkers at the patient (slide) level (e.g., microsatellite instable vs microsatellite stable for MSI, high mutation density vs low mutation density for hypermutation, chromosomal instability vs genomic stability for chromosomal instability, CIMP-high vs CIMP-low for CIMP, mutant vs. wild type for BRAF, and TP53 mutation). The linear project layers of the saved pre-trained Swin-T model for tissue classification were replaced by new linear layers to accommodate the prediction for the binary classification for the CRC markers. The new model was fine-tuned for 20 epochs in the binary classification task for each binary CRC biomarker by using a decaying learning rate policy (i.e., the learning rate at the nth epoch is 0.0001/n). Similarly, Adam optimizer and cross entropy loss function with weight were used during fine-tuning. The average value of the predictive scores of all tiles of each WSI is calculated as the score of the corresponding molecular status of the whole-slide image.

**Experimental setup**

In total, three experiments were performed to evaluate the performance of our Swin-T models. First, we compared the predictive performance of the Swin-T models with that of state-of-the-art models for predicting 6 CRC biomarkers (hypermutation, microsatellite instability, chromosomal instability, CIMP, BRAF, and TP53 mutation) from two recent publications using intra-study cross validation. For all 6 biomarkers, we used the exact same training-to-test dataset spilt of the TCGA-CRC-DX dataset for a four-fold cross-validation that was used and published by Kather and colleagues or/and Bilal and colleagues. The match of the split of the TCGA-CRC-DX cohort was to facilitate the comparison of model performance with previous publications. In cross-validation, one fold of the training set was used as the validation set to select best performing model, which was saved for testing on the unseen test fold.
Second, the predictive performance of the Swin-T models was compared with the state-of-the-art models from recent publications using cross-study external validation for predicting microsatellite instability and *BRAF* mutation\(^{11,14}\). For the external-validation experiments, the pre-trained Swin Transformer model was fine-tuned using the MCO cohort to develop the models for predicting MSI status and BRAF mutation. In addition, a model for predicting CIMP status was also developed as the molecular data for these 3 biomarkers are available in both TCGA and MCO cohorts). The fine-tuned models were tested externally on the unseen TCGA-CRC-DX cohort in this exercise to facilitate the comparison with publications in which the external validation was performed on the same cohort.

Finally, to understand the impact of sample size of the training data on the performance of the Swin-T models, we randomly selected 25%, 50%, and 75% of the MCO data, and trained additional models for prediction of MSI status using the same training strategy as that used during the training with all MCO data. Similarly, the external-validation performance on the TCGA-CRC-DX cohort were compared at the different sample sizes for training data.

**Statistical analyses**

Predictive performance of deep-learning models was evaluated using different statistical metrics. For the intra-study cross-validation, the mean values of area under the receiver operating curve (AUROC) and the area under the precision-recall curve (AUPRC) were used by averaging four-fold results. AUPRC was compared to account for the imbalanced data commonly seen in clinical data. Standard deviations were calculated. For external validation, AUROC and AUPRC were also utilized to compare models. Bootstrap method (1000×) was applied to calculate 95% confidence intervals (CI) for external validation AUROC and AUPRC. To evaluate the feasibility of the models as diagnostic tools, different classification thresholds were set to calculate a variety
of statistical metrics to evaluate diagnostic performance based the external-validation experiments. The thresholds included cutoff at fixed 95% sensitivity as well as cutoff at fixed values of 0.25, 0.50 and 0.75. Sensitivity, specificity, positive predictive value, negative predictive value, true negative fraction, false negative fraction, and F1-score were used to evaluate the diagnostic values of the AI models. In addition, the interpretability of the Swin-T models was explored using the visualization technology with the python package pytorch_grad_cam (https://github.com/jacobgil/pytorch-grad-cam). The Grad-CAM method\textsuperscript{24} was applied to visualize the activation feature map of Swin-T model and interpret the model outcome.

**Results**

1. **Swin-T provides excellent predictive performance**

   **Intra-study cross validation using TCGA-CRC-DX dataset**

   In this experiment, a total of six molecular biomarkers (i.e., MSI, hypermutation, CIMP, CING, BRAF mutation, and TP53 mutation) were predicted using Swin-T. Kather et. al.\textsuperscript{6} and Bilal et al.\textsuperscript{4} used a TCGA-CRC-DX dataset to develop and evaluate their deep-learning models for MSI status and other key biomarkers for CRC via intra-study cross validations. To facilitate comparisons with their existing models, we used the same patient cohort, and same 4-fold splits of the TCGA dataset which published by Bilal et al.\textsuperscript{4}.

   For the prediction of high microsatellite instability status, our Swin-T achieved a mean validation AUROC of 0.91 ± 0.03 (mean ± SD), which represents approximately 6% and 23% improvement over recently published AUROC values on the same dataset, i.e., 0.86 from Bilal et
al.\textsuperscript{4} and 0.74 from Kather et al.\textsuperscript{12}, respectively (\textbf{Table1}). For prediction of hypermutation status, Swin-T also outperformed the models developed by Kather et. al.\textsuperscript{6} and Bilal et al.\textsuperscript{4} on the same dataset, and significantly improves the predictive performance, i.e., the AUROC based on Swin-T was 0.85±0.02, compared to 0.81 and 0.71 reported by Kather et al. Bilal et al., respectively. Swin-T showed similar performance to Bilal et al.’s method for predicting TP53 mutation status (AUROC = 0.73), but was significantly higher than that was obtained in Kather et. al.\textsuperscript{6}(AUROC = 0.64). In addition, although our mean cross-validation AUROC values for predicting chromosomal instability (vs. genomic stability), \textit{BRAF} mutation status, and CIMP high status were slightly lower than Bilal et al.’s methods, the difference was only 1-2%.

Overall, for a fixed sample size for training (the same TCGA-CRC-DX dataset), Swin-T provided significant improvement for predicting MSI status and hypermutation status in intra-study cross validation over published models. Also, Swin-T achieved SOTA performance for predicting \textit{TP53} mutation. Furthermore, Swin-T models provided similar or higher AUPRC for MSI, hypermutation, CIMP, and \textit{BRAF} mutation status compared to the models reported by Bilal et al.\textsuperscript{4}, suggesting that the Swin-T models better predict positive (MSI-high, high mutation density, CIMP high, and \textit{BRAF} mutants) with lower false positive rate.

\textbf{Cross-study external validation using TCGA-CRC-DX dataset}

Generalizability of a model is often evaluated by cross-study external validation. The TCGA-CRC-DX dataset has been used in external validation for multiple AI models for predicting MSI-status. In this experiment, we trained the model using the MCO dataset (N = 1065), and then externally validated the model using the TCGA-CRC-DX dataset to compare the model performance. The Swin-T model yielded an excellent mean external-validation AUROC
of 0.904 (95% confidence interval (CI): 0.849-0.952; Table 2). In comparison, Echle et al.\textsuperscript{11} trained a CNN model using a combined dataset from multiple large international studies (N = 7917) and achieved a similar mean external-validation AUROC of 0.91 (95% CI: 0.87 – 0.95). The model trained using ShuffleNet and similar size of training data (N ranging from approx. 1000 – 2000) only produced AUROC values from 0.72 to 0.77 (Table 2)\textsuperscript{12}. Therefore, Swin-T achieved similar, SOTA generalizability in external validation compared to the most recently published model trained on a large, pooled dataset.

Swin-T also showed similar, SOTA performance to previously published methods for predicting \textit{BRAF} mutation status (0·80 [0.74 – 0.87] vs 0·81 [0.75 – 0.86]) in external validation using the TCGA-CRC-DX dataset. It is worth noting that Swin-T produced substantially better AUPRC values compared to previous publications for predicting both MSI-status (0.66 vs. 0.62) and \textit{BRAF} mutation (0.35 vs 0.33).

Furthermore, the Swin-T architecture demonstrated great potential minimizing overfitting that often observed in deep-learning modeling and produced very similar predictive performance between training dataset and external validation dataset. In the four-fold cross-validation experiment of the MCO dataset for predicting MSI-status, Swin-T achieved a mean AUC value of 0.926±0.055 compared to 0·904 (95%CI: 0.849-0.952) in external validation dataset (TCGA-CRC-DX). Similar pattern was observed for \textit{BRAF} (0.88 vs. 0.80) and CIMP (0.766 vs. 0.759) (Figure 3).

2. Swin-T models as diagnostic tools

Based on the external validation using the TCGA-CRC-DX cohort, we also evaluate the feasibility of using Swin-T models as diagnostic tools for MSI/dMMR status, \textit{BRAF} mutation,
CIMP status based on routine digitized H&E-stained tissue slides of CRC. Computer-based AI systems are often positioned as pre-screening tools before gold-standard confirmatory tests\textsuperscript{11}. Therefore, the clinical utility of these pre-screening tools is mainly to minimize the false negative predictions but exclude as many true negative samples as possible from the subsequent confirmatory test runs.

**Prescreening for MSI status**

For diagnostic purposes, a cutoff is needed to decide on the diagnostic outcome. Table 3 shows that, for a cutoff that can provide 95% sensitivity for detecting MSI-High patients (cutoff = 0.16), the negative predictive value was 98% and the false negative fraction was only 0.7%. Meanwhile, the true negative fraction was 34% with this cutoff, which means that 34% of patients can be safely exclude from the confirmatory tests in clinical settings. Similar results were reported by Echle et al.\textsuperscript{11} for MSI detection. When using a fixed cutoff of 0.25, the sensitivity reduced slightly to 92%. The NPV remained almost the same at 98%, while FNF slightly increase to 1.2%. However, with the cutoff of 0.25, 55.4% of patients can be excluded from the confirmatory tests. These results confirmed the previous report that AI models can serve as prescreening purpose for MSI status\textsuperscript{8,11,12}.

**Prescreening for BRAF mutation**

For the Swin-T model for *BRAF* mutation, the cutoff for 95% sensitivity was 0.17. At such a cutoff, the NPV was 96%, the FNF was 0.4%, while the TNF 10%, suggesting that only 10% *BRAF* WT patients would be safely excluded from the gold-standard confirmatory testing for *BRAF* mutation. However, if we use a fixed cutoff of 0.25, 27.2% patients could be correctly
determined as BRAF WT, whereas the false negatives (patients incorrectly determined as BRAF WT) remained very low (0.8%). 96% predicted BRAF WT were actually BRAF WT at the cutoff of 0.25. The Swin-T model for BRAF mutation showed the potential as a prescreening AI diagnostic tool for BRAF mutation.

**Prescreening for CIMP status**

A cutoff of 0.1 can provide 95% sensitivity for predicting CIMP status. At this cutoff, 1.3% CIMP high would be incorrectly identified as CIMP-low, while 16.2% true CIMP-low can be excluded from the subsequent confirmatory molecular testing. However, when increasing the cutoff to 0.25, the false negatives could substantially increase to approx. 5%. Therefore, the performance for the current Swin-T model for CIMP may not be optimal as a diagnostic tool.

**3. Swin-T exhibits robustness with small training data**

As expected, the larger the training datasets, the higher predictive performance can be achieved (Figure 5). However, the Swin-T model maintained the cross-study AUROC at 0.864 (on external validation dataset TCGA-CRC-DX) for predicting microsatellite instability when only 50% of MCO data (N = ~ 500) were used in training. In the scenario using only 25% MCO dataset for training (N = ~250), the Swin-T model still managed to produce an impressive AUROC value of 0.806 for predicting MSI status.

**4. Swin-T models improve interpretability**

Tile-level interpretation of deep-learning models for MSI has been popular and often provided in current literature\textsuperscript{1,3,8,12}. We also visualized the highest scoring tiles of the patients
with the highest WSI scores (**Figure 6a**). For Swin-T, we could draw a cam map for a tile (**Figure 6b**), which is a heatmap to highlight the areas (cells) that have greater attention weights within a high-resolution tile. The brighter the color for an area in the heatmap of a tile, the higher attention scores are assigned to this area by the model. **Figure 6b** shows that, within the tiles of MSI samples, the model focuses on the microenvironment around the tumor cells rather than the tumor cells themselves, including mucus, infiltrating lymphocytes, tumor stroma, etc. However, within the tiles predicted from the MSS sample, the model mainly focused on the tumor tissues. These findings were consistent with the pathological characteristics of MSI-high or MSS samples\(^{25,26}\), suggesting that coupling Swin-T with Grad-CAM algorithm\(^{24}\) could further improve the interpretability of the model and detect the areas or cells within tiles that have great potential aid pathologists to identify novel image features associated with classification of a biomarker.

**Discussion**

In this study, we developed a novel deep-learning framework based on a Swin-T backbone network for predicting MSI status and other key biomarkers for CRC. The Swin-T backbone represents the most advanced, start-of-the-art vision transformer network architecture and have achieved outstanding performance in many computer-vision tasks, outperforming many popular, de-facto standard networks such as ResNet and EfficientNet as well as early version of vision transformers (ViT)\(^{17,18}\). It has been proved that Swin Transformer can replace the classic CNN architecture and become a common backbone in the field of computer vision\(^{18}\). However, despite having achieved great success on common computer vision tasks, to our knowledge, our work represents the first attempt to evaluate Swin-T’s performance in digital pathology and as a
backbone network for further improving the predictive performance of MSI and biomarkers for molecular pathways in CRC.

We demonstrated that the novel Swin-T based backbone networks have great utility in digital pathology as well and can improve the predictive performance for microsatellite instability and other key biomarkers in colorectal cancer tumors. To facilitate the comparison with previously published models, the exact same dataset (TCGA-CRC-DX) and training-to-test spilt of the dataset from previous publications was used. In the intra-study cross-validation experiment, Swin-T substantially outperformed models by Bilal and colleagues\(^4\) and Kather colleagues\(^12\) for prediction of microsatellite instability and hypermutation status. In addition, Swin-T achieved similar SOTA performance for predicting \textit{TP53} mutation status compared to the work by Bilal and colleagues\(^4\). Similar mean cross-validation AUROC values were also obtained for predicting chromosomal instability, \textit{BRAF} mutation status, and CIMP high status compared to the current literature\(^4\). What’s more, Swin-T models also exhibited similar or higher AUPRC for MSI, hypermutation, CIMP, and BRAF mutation status compared to the previously published state-of-the-art computational algorithms\(^4\), indicative of greater power for handling imbalanced data which is often seen in clinical studies.

It is well known that deep learning models perform better with more training data available. This phenomenon has been observed for prediction models developed for MSI/dMMR status in CRC\(^12\). Most recently, Echle et al. trained a model using pooled data from nine patient cohorts of 8343 patients across different countries and ethnicities and achieved the state-of-the-art (SOTA) external prediction performance with an AUC of 0.91 using the TCGA-CRC-DX cohort as the external validation dataset\(^11\). However, with smaller training data (QUASAR: N=1016; DACHS: N=2013; NLCS: N=2197), Echle et al. (the same research group) only obtained an AUROC of
0.72 – 0.77 with the same unseen external validation cohort. Swin-T showed excellent generalizability in cross-study external validation using the same TCGA-CRC-DX dataset, and delivered a SOTA AUROC of 0.904 using the relatively smaller training data (MCO, N = 1065), similar to what was achieved by Echle and colleagues using ~8000 samples (ResNet18). Our additional experiment also showed that Swin-T is extremely efficient using small training datasets. Using only ~250 samples for training, the Swin-T model still managed to produce better predictive performance than Echle and colleagues’ models using ShuffleNet and training data of 1000 – 2000 samples. These results suggest that our MSI model based on Swin-T may be 5 – 10 times more efficient than the current state-of-the-art MSI algorithms based on ResNet18 and ShuffleNet.

Biomarker testing plays a critical role in treatment selection for patients with CRC. Importantly, immunotherapies such as pembrolizumab and nivolumab have been approved by health authorities to treat CRC patients with MSI-High. The current clinical gold-standard testing for MSI is based on immunohistochemistry, which has a sensitivity of 94% and a specificity of 88% [ref]. The motivation to develop AI-based models is mainly to replace the current lab-based testing, reduce turnaround time, and save cost. Unfortunately, so far, no digital AI models for MSI can consistently achieve this performance threshold, including the most recent model developed by Echle and colleagues. Therefore, it is proposed to implement current state-of-the-art MSI models as a pre-screening test, which is primarily to exclude and reduce the samples before the subsequent conventional IHC testing. Therefore, for the clinical utilization, achieving a high rule-out fraction (true-negative fraction) and a low false-negative fraction is critical. Our Swin-T model for MSI status showed a similar diagnostic performance compared to the model developed by Echle and colleagues that was trained with approximately
8000 CRC patients. In addition, the Swin-T model for BRAF mutation status also showed promise as a pre-screening test although further improvement might be needed. These results demonstrated the potential of this Swin-T-based AI system to be an important component in a cascading diagnostic workflow (pre-screening + gold-standard testing) for MSI status and BRAF mutation status, which are important for patient selection in clinical trials and treatment guidance for immune checkpoint inhibitors and combinations of BRAF inhibitors/anti-epidermal growth factor receptor therapies, respectively.

Reference

1. Kather JN, Pearson AT, Halama N, Jger D, Luedde T. Deep learning can predict microsatellite instability directly from histology in gastrointestinal cancer. *Nature medicine*. 2019;25(7).
2. Schmauch B, Romagnoni A, Pronier E, et al. A deep learning model to predict RNA-Seq expression of tumours from whole slide images. *Nature communications*. 2020;11(1):1-15.
3. Yamashita R, Long J, Longacre T, et al. Deep learning model for the prediction of microsatellite instability in colorectal cancer: a diagnostic study. *The Lancet Oncology*. 2021;22(1):132-141.
4. A M, Shan E, Aaa B, 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 - ScienceDirect. *Lancet Digit Health*. 2021.
5. Fu Y, Jung AW, Torne RV, et al. Pan-cancer computational histopathology reveals mutations, tumor composition and prognosis. *Nature Cancer*. 2020;1(8):800-810.
6. Kather JN, Heij LR, Grabsch HI, et al. Pan-cancer image-based detection of clinically actionable genetic alterations. *Nature Cancer*. 2020;1(8):789-799.
7. Sirinukunwattana K, Domingo E, Richman SD, et al. Image-based consensus molecular subtype (imCMS) classification of colorectal cancer using deep learning. *Gut*. 2021;70(3):544-554.
8. 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*. 2021:100008.
9. Marcus L, Lemery SJ, Keegan P, Pazdur R. FDA Approval Summary: Pembrolizumab for the Treatment of Microsatellite Instability-High Solid Tumors. *Clinical Cancer Research*. 2019;25(13):3753-3758.
10. André T, Shiu K-K, Kim TW, et al. Pembrolizumab in microsatellite-instability-high advanced colorectal cancer. *New England Journal of Medicine*. 2020;383(23):2207-2218.
11. Echle A, Laleh NG, 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.
12. Echle A, Grabsch HI, Quirke P, et al. Clinical-grade detection of microsatellite instability in colorectal tumors by deep learning. *Gastroenterology*. 2020;159(4):1406-1416. e1411.
13. Lee SH, Song IH, Jang HI. Feasibility of deep learning–based fully automated classification of microsatellite instability in tissue slides of colorectal cancer. *International Journal of Cancer*. 2021;149(3):728-740.
14. Laleh NG, Muti HS, Loeffler CML, et al. Benchmarking weakly-supervised deep learning pipelines for whole slide classification in computational pathology. *Medical image analysis*. 2022;79:102474.
15. 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. *Medical Image Analysis.* 2022;79:102464.

16. Cao R, Yang F, Ma S-C, et al. Development and interpretation of a pathomics-based model for the prediction of microsatellite instability in Colorectal Cancer. *Theranostics.* 2020;10(24):11080.

17. Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:11929.* 2020.

18. Liu Z, Lin Y, Cao Y, et al. Swin transformer: Hierarchical vision transformer using shifted windows. Paper presented at: Proceedings of the IEEE/CVF International Conference on Computer Vision2021.

19. Schrammen PL, Ghaffari Laleh N, Echle A, et al. Weakly supervised annotation-free cancer detection and prediction of genotype in routine histopathology. *The Journal of pathology.* 2022;256(1):50-60.

20. Guo B, Li X, Yang M, Zhang H, Xu XS. A robust and lightweight deep attention multiple instance learning algorithm for predicting genetic alterations. *arXiv preprint arXiv:00455.* 2022.

21. Macenko M, Niethammer M, Marron JS, et al. A method for normalizing histology slides for quantitative analysis. Paper presented at: 2009 IEEE international symposium on biomedical imaging: from nano to macro2009.

22. Kather JN, Krisam J, Charoentong P, et al. Predicting survival from colorectal cancer histology slides using deep learning: A retrospective multicenter study. *PLoS medicine.* 2019;16(1):e1002730.

23. Li X, Jonnagaddala J, Yang S, Zhang H, Xu XS. A Retrospective Analysis using Deep-Learning Models for Prediction of Survival Outcome and Benefit of Adjuvant Chemotherapy in Stage II/III Colorectal Cancer. *arXiv preprint arXiv:211103532.* 2021.

24. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-cam: Visual explanations from deep networks via gradient-based localization. Paper presented at: Proceedings of the IEEE international conference on computer vision2017.

25. Greenson JK, Huang S-C, Herron C, et al. Pathologic predictors of microsatellite instability in colorectal cancer. *The American journal of surgical pathology.* 2009;33(1):126.

26. Jenkins MA, Hayashi S, O’shea A-M, et al. Pathology features in Bethesda guidelines predict colorectal cancer microsatellite instability: a population-based study. *Gastroenterology.* 2007;133(1):48-56.
Table 1. Comparison of predictive performance of Swin-T for key CRC biomarkers with published models using intra-study cross validation on TCGA-CRC-DX

| Biomarker                                      | AUROC        | AUPRC        |
|------------------------------------------------|--------------|--------------|
|                                                | Swin-T       | Kather et al.\(^6\) | Bilal et al.\(^4\) (IDaRS) | Swin-T | Bilal et al.\(^4\) (IDaRS) |
| Four-fold cross-validation in the TCGA-CRC-DX cohort |              |              |                          |        |                            |
| Microsatellite instability vs stability         | 0.91± 0.02   | 0.74         | 0.86± 0.03               | 0.66± 0.09 | 0.62± 0.10 |
| High vs low mutation density                    | 0.85± 0.03   | 0.71         | 0.81±0.04                | 0.58 ± 0.05 | 0.57± 0.09 |
| Chromosomal instability vs genomic stability    | 0.81± 0.04   | 0.73         | 0.83±0.02                | 0.90± 0.03 | 0.92± 0.01 |
| CIMP-high vs CIMP-low                           | 0.77± 0.06   | ...          | 0.79±0.05                | 0.60± 0.15 | 0.51± 0.05 |
| BRAF                                           | 0.77± 0.02   | 0.66         | 0.79±0.01                | 0.35± 0.11 | 0.33± 0.05 |
| TP53                                           | **0.73± 0.02** | 0.64        | **0.73± 0.02**           | 0.75± 0.02 | **0.78± 0.04** |
Table 2. Comparison of predictive performance of MCO-trained Swin-T for MSI status and BRAF mutation with published models using external validation on TCGA-CRC-DX.

| Network | Training Dataset | Number of training samples | AUROC (95% CI) | AUPRC (95% CI) |
|---------|------------------|-----------------------------|----------------|----------------|
| **MSI** |                  |                             |                |                |
| Swin-T (Ours) | MCO     | 1065                        | 0.90 (0.85-0.95) | 0.718 (0.605-0.820) |
| VIT\(^{14}\) (2022) | DACHS   | 2069                        | 0.89 (0.83-0.93) | 0.672 (0.558-0.769) |
| ResNet18\(^{11}\) (2022) | Pooled International datasets | 7917                        | 0.91 (0.87-0.95) | … |
| ShuffleNet\(^{12}\) (2020) | QUASAR | 1016                        | 0.76 (0.70-0.79) | … |
|                  | DACHS   | 2013                        | 0.77 (0.73-0.79) | … |
|                  | NLCS    | 2197                        | 0.72 (0.71-0.78) | … |
| **BRAF** |                  |                             |                |                |
| Swin-T (Ours) | MCO     | 1026                        | 0.80 (0.74-0.87) | 0.392 (0.279-0.541) |
| EfficientNet\(^{14}\) (2022) | DACHS   | 2069                        | 0.81 (0.75-0.86) | 0.36 (0.253-0.487) |
Table 3. Statistics results using different thresholds of external validation of predictions for the MSI, BRAF and CIMP status in TCGA-CRC-DX cohort.

| Biomarker | Threshold | Sensitivity | Specificity | PPV  | NPV  | TNF  | FNF  | F1 Score |
|-----------|-----------|-------------|-------------|------|------|------|-------|----------|
| MSI       | 0.16      | 0.95        | 0.402       | 0.209| 0.980| 0.340| 0.007 | 0.343    |
|           | 0.25      | 0.918       | 0.647       | 0.304| 0.979| 0.554| 0.012 | 0.457    |
|           | 0.5       | 0.721       | 0.939       | 0.667| 0.953| 0.804| 0.040 | 0.693    |
|           | 0.75      | 0.213       | 0.997       | 0.928| 0.883| 0.853| 0.113 | 0.346    |
| BRAF      | 0.17      | 0.95        | 0.114       | 0.123| 0.959| 0.095| 0.004 | 0.218    |
|           | 0.25      | 0.930       | 0.307       | 0.148| 0.971| 0.272| 0.008 | 0.255    |
|           | 0.5       | 0.649       | 0.827       | 0.327| 0.948| 0.732| 0.040 | 0.435    |
|           | 0.75      | 0.263       | 0.977       | 0.6   | 0.911| 0.865| 0.085 | 0.366    |
| CIMP      | 0.10      | 0.95        | 0.201       | 0.263| 0.927| 0.162| 0.013 | 0.411    |
|           | 0.25      | 0.796       | 0.536       | 0.339| 0.898| 0.413| 0.047 | 0.475    |
|           | 0.5       | 0.556       | 0.845       | 0.517| 0.864| 0.651| 0.102 | 0.536    |
|           | 0.75      | 0.278       | 0.945       | 0.6   | 0.814| 0.728| 0.166 | 0.380    |

Statistics describe the different thresholds when the network is trained on MCO cohorts and tested on TCGA-CRC-DX cohort. PPV: Positive Predictive Value. NPV: Negative Predictive Value. TNF: True-Negative Fraction (Rule-out). FNF: False-Negative Fraction.
**Figures**

**Figure 1:** The workflow of the data preprocessing and the training process of the Deep-Learning model. (a) Tiles images of NCT-CRC-HE-100K are downloaded from the publicly available website (https://zenodo.org/record/1214456) to pre-train a tissue classifier based on Swin-T. The classifier is proven to have excellent performance of classify tissues in an external dataset: CRC-VAL-HE-7K. (b) Whole-Slide images in the SVS format of the MCO dataset and TCGA dataset are preprocessed to tessellate into non-overlapping patches with a size of 512 × 512 pixels. These tiles are then resized to the smaller 224×224 pixels tiles and color normalized. Then the pre-trained tissue classifier in (a) is applied to select tumor tiles. (c) For each patient, up to 500 tiles are randomly sampled for subsequent experiments. The pre-trained model is then re-trained to predict outcome the molecular pathway of each tile then the results of the tiles are pooled on patient level. The performance of the models is evaluated both in 4-fold cross-validation within-cohort and in external cohort.

**Figure 2:** Results of four-fold cross-validation of Swin-T based prediction of colorectal cancer pathways in the TCGA-CRC-DX cohort. AUROC plots for prediction of hypermutation (HM), microsatellite instability (MSI), chromosomal instability (CING), CpG island methylator phenotype (CIMP), BRAF mutation status and TP53 mutation status. The True Positive Rate represents sensitivity and the False Positive Rate represents 1–specificity. The red shaded areas represent the SD. The value in the lower right of each plot represents mean AUROC± SD.

**Figure 3:** Results of intra-cohort four-fold cross-validation in MCO cohort and inter-cohort external validation in TCGA-CRC-DX cohort. (a) AUROC plots for four-fold cross-validation in MCO cohort of microsatellite instability, BRAF mutation status, CpG island methylator phenotype. The red shaded areas represent the SD. The value in the lower right of each plot represents mean AUROC ± SD. (b) AUROC plots for inter-cohort external validation in TCGA-CRC-DX cohort: microsatellite instability, BRAF mutation status, CpG island methylator phenotype. The red shaded areas represent the 95% confidence interval (CI), which calculated by 1000× bootstrap. The value in the lower right of each plot represents mean AUROC (95% CI).

**Figure 4. Test statistics for the pre-screening tool.** Test performance of MSI status, BRAF mutation and CIMP status in the TCGA-CRC-DX cohorts displayed as patients classified true/false positive/negative by the Swin-T model based on 95% sensitivity thresholds, fixed thresholds (0.25, 0.5 and 0.75).

**Figure 5. The visualization and interpretability of Swin-T model in predicting MSI status.** (a) Tile-level interpretation. we plot the top accurately predicted tiles from the top accurately scoring patients, i.e., for positive specimens (MSI-High), visualize the highest scoring tiles of the patients with the highest scores. Similar operation was done for negative specimens (MSS), but the lowest scoring tiles are shown. 50 tiles from the MCO and TCGA datasets are visualized, separately. (b) Cell-level interpretation. we could draw a cam map for a tile using Grad-CAM algorithm, which is a heatmap to highlight the areas (cells) that have greater attention weights within a high-resolution tile. The brighter the color for an area in the heatmap of a tile, the higher
attention scores were assigned to this area by the model.

**Figure 6. Results of Swin-T for prediction of MSI status with an increasing small sample size.** The plot shows that the test performance in external TCGA-CRC-DX dataset of Swin-T, when it is trained with an increasing MCO training dataset (25%, 50%, 75%, and 100%). The bar plots show AUROC values and the error bars represent the 95% confidence intervals.
Figure 1.

(a) NCT-CRC-HE-100K dataset → Pre-train model for tissue classification task → Test in CRC-HE-7K dataset

(b) MCO dataset → TCGA dataset → Tessellate into tiles → Color normalization and select tumor tiles

(c) Collect up to 500 tiles per patient → Train pre-trained model to predict each tile → Pool on patient level

- 4-fold cross-validation within-cohort
- Test in external cohort
Figure 2.
Figure 3.

**Intra-study cross validation in MCO**

- MSI
  - True Positive Rate vs False Positive Rate
  - Mean AUROC: 0.926 ± 0.055

- BRAF
  - True Positive Rate vs False Positive Rate
  - Mean AUROC: 0.877 ± 0.019

- CIMP
  - True Positive Rate vs False Positive Rate
  - Mean AUROC: 0.765 ± 0.051

**Cross-study external validation in TCGA**

- MSI
  - True Positive Rate vs False Positive Rate
  - AUROC: 0.904 (0.849-0.952)

- BRAF
  - True Positive Rate vs False Positive Rate
  - AUROC: 0.800 (0.729-0.861)

- CIMP
  - True Positive Rate vs False Positive Rate
  - AUROC: 0.759 (0.676-0.836)
Figure 4

| TCGA MSI | TCGA BRAF | TCGA CIMP |
|---------|-----------|-----------|
| True Positives | False Positives | False Negatives | True Negatives |
| 56 | 44 | 13 | 58 | 128 | 22 | 1 | 58 |
| 5 | 17 | 48 | 3 | 235 | 341 | 362 | 144 |
| 305 | 76 | 10 | 55 | 135 | 364 | 430 | 47 |
| 3 | 2 | 54 | 2 | 84 | 28 | 153 | 38 |
Figure 5.
Figure 6

Test AUCs in TCGA-CRC-DX

Samples of training set

| Samples of training set | Test AUCs |
|-------------------------|-----------|
| 266 (25%)               | 0.806     |
| 532 (50%)               | 0.864     |
| 798 (75%)               | 0.877     |
| 1065 (100%)             | 0.904     |

AUROC
95% CI
Supplement Figure 1: Results (AUPRC) of four-fold cross-validation of Swin ViT-based prediction of colorectal cancer pathways in the TCGA-CRC-DX cohort. AUPRC plots for prediction of (a) hypermutation, (b) microsatellite instability, (c) chromosomal instability, (d) CpG island methylator phenotype, (e) BRAF mutation status and (f) TP53 mutation status. The red shaded areas represent the SD. The value in the lower right of each plot represents mean AUPRC ± SD.