Self-supervised learning improves dMMR/MSI detection from histology slides across multiple cancers

Editor: -

Abstract

Microsatellite instability (MSI) is a tumor phenotype whose diagnosis largely impacts patient care in colorectal cancers (CRC), and is associated with response to immunotherapy in all solid tumors. Deep learning models detecting MSI tumors directly from H&E stained slides have shown promise in improving diagnosis of MSI patients. Prior deep learning models for MSI detection have relied on neural networks pretrained on ImageNet dataset, which does not contain any medical image. In this study, we leverage recent advances in self-supervised learning by training neural networks on histology images from the TCGA dataset using MoCo V2. We show that these networks consistently outperform their counterparts pretrained using ImageNet and obtain state-of-the-art results for MSI detection with AUCs of 0.92 and 0.83 for CRC and gastric tumors, respectively. These models generalize well on an external CRC cohort (0.97 AUC on PAIP) and improve transfer from one organ to another. Finally we show that predictive image regions exhibit meaningful histological patterns, and that the use of MoCo features highlighted more relevant patterns according to an expert pathologist.

Keywords: Self-supervised learning, Microsatellite Instability (MSI)

1. Introduction

Microsatellite Instability (MSI) is a frequent tumor phenotype characterized by an abnormal repetition of short DNA motifs caused by a deficiency of the DNA mismatch repair system (MMR). MMR deficient tumors (dMMR) result from defects in the major MMR genes, namely *MLH1, MSH2, MSH6, PMS2*. These defects arise either sporadically or as a hereditary condition named Lynch syndrome (LS), predisposing patients to develop cancers in several organs.

Recent studies have shown that immune checkpoint blockade therapy has a promising response in dMMR/MSI cancers regardless of the tissue of origin [Le et al. (2017)]. In 2017, this genomic instability phenotype became the first pan-cancer biomarker approved by the US FDA, allowing the use of pembrolizumab (Keytruda) for patients with dMMR/MSI solid tumors at the metastatic stage [Prasad et al. (2018)].

As of today, systematic MSI screening is only recommended for colorectal cancer (CRC) and endometrial cancer [Svrcek et al. (2019)] where the prevalence is relatively high (10% to 20%), principally to detect LS patients and provide them with adequate follow-up. In early stages of CRC, MSI tumors are associated with good prognosis and resistance to chemotherapy [Sargent et al. (2010)], making the diagnosis of this phenotype all the more essential for patient care and therapeutic decision. dMMR/MSI diagnosis is traditionally done using immunohistochemistry (IHC), polymerase chain reaction (PCR) assays, or next

generation sequencing. Those methods can be time-consuming, expensive, and rely on specific expertise which may not be available in every center.

Deep learning based MSI classifiers using H&E stained digital images offer a new alternative for a broader and more efficient screening [Echle et al. (2020)]. In CRC, previous work suggests that the use of such models as pre-screening tools could eventually replace IHC and PCR for a subset of tumors classified as microsatellite stable (MSS) or unstable (MSI) with a high probability [Kacew et al. (2021)]. In other locations where MSI prevalence is lower and screening not done as routine practice, predictive models of MSI status from WSI could be used as efficient pre-screening tools.

In this work, we leverage recent advances in self-supervised learning (SSL) on images. We show that SSL permits to reach state-of-the-art results on colorectal and gastric cancer cohorts from The Cancer Genome Atlas (TCGA), generalizes well on an unseen colorectal cohort (PAIP), and could pave the way for classifiers on locations with low MSI prevalence.

2. Related Work

Expert models Several histology patterns on H&E images have been reported to correlate with MSI, such as tumor-infiltrating lymphocytes, lack of dirty necrosis or poor differentiation. A series of models based on clinico-pathological features have been developed [Greenson et al. (2009); Jenkins et al. (2007); Hyde et al. (2010); Fujiyoshi et al. (2017); Román et al. (2010)] and reported ROC-AUC performances ranging from 0.85 to 0.92 in various cohorts of patients with CRC. These methods however require time-consuming annotations from expert pathologists and are prone to inter-rater variability.

Deep learning models In a seminal publication, Kather et al. (2019) proved the feasibility to determine the dMMR/MSI status from H&E stained whole slide images (WSI) using deep learning. They trained a first ResNet [He et al. (2016)] to segment tumor regions on WSI, and a second one to predict MSI/MSS status in each tumor tile. Each ResNet was pretrained on ImageNet and the last 10 layers were fine-tuned. Models were trained and validated on different TCGA cohorts and obtained respectively AUCs of 0.77, 0.81 and 0.75 on Colorectal, Gastric and Endometrial formalin-fixed paraffin-embedded (FFPE) datasets.

In a larger scale study focusing on CRC only [Echle et al. (2020)], the same team later trained a model on $n = 6406$ patients, reaching 0.96 AUC (95% CI 0.93–0.98) on an external dataset of $n = 771$ patients. Tumor tissues were manually outlined by pathologists.

Since then, different works based on deep learning methods have been published [Zhang et al. (2018); Cao et al. (2020); Hong et al. (2020); Yamashita et al. (2021); Bilal et al. (2021); Lee et al. (2021)] and are reviewed in Hildebrand et al. (2021). The vast majority rely on networks pretrained on the ImageNet dataset [Deng et al. (2009)] and only the last layers are re-trained or fine-tuned. Most of these papers also rely on tumor segmentation as a first step in their models (either by a pathologist or by a deep learning model).

Self-Supervised learning Over the past few years, rapid progress has been made in the field of SSL using contrastive learning or self-distillation strategies: simCLR [Chen et al. (2020a,b)] , MoCo [He et al. (2020); Chen et al. (2020c, 2021)], BYOL [Grill et al. (2020)], achieving impressive performances on ImageNet without using any labels. Such models

have also been shown to outperform supervised models in transfer learning tasks [Chen et al. (2020a); Caron et al. (2021); Li et al. (2021b)].

These advances are of particular interest in medical imaging applications where labeled datasets are hard to collect, and especially in histology where each WSI contains thousands of unlabeled images. There is growing evidence that SSL is a powerful method to obtain relevant features for various prediction tasks from histology images [Dehaene et al. (2020); Lu et al. (2019); Li et al. (2021a); Koohbanani et al. (2021); Gildenblat and Klaiman (2019); Abbet et al. (2021); Srinidhi et al. (2021)]. In this work, we show that self-supervision can be efficiently used to detect dMMR/MSI tumors from histology slides, and outperform ImageNet pretrained models across and between several tumors.

3. Methods

3.1 Proposed pipeline

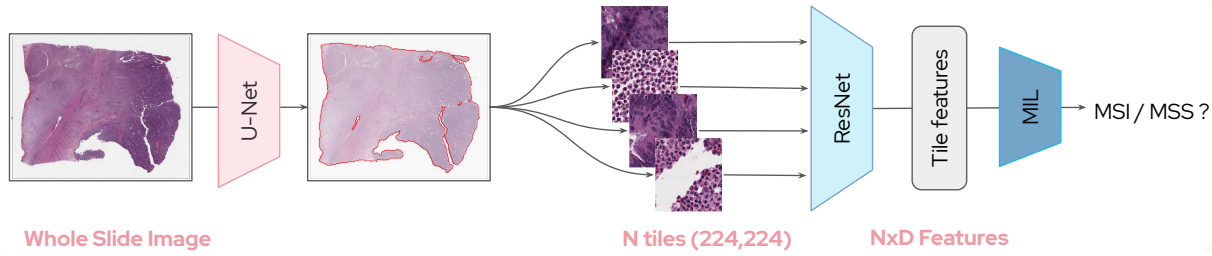

Figure 1: Overview of the proposed pipeline.

First, a U-Net neural network [Ronneberger et al. (2015)] is used to segment tissue on the input WSI and discard the background, as well as artifacts. Second, segmented tissue is divided into N (typically between 10,000 and 30,000) smaller images called tiles. Each tile has a fixed shape of 224×224 pixels (resolution of 0.5 micron per pixel). Third, the N tiles are embedded into feature vectors of shape D using a pretrained convolutional neural network. Fourth, the $N \times D$ features are aggregated using a multiple instance learning model. This final model is the only one trained using MSI/MSS labels.

In this study, we benchmarked 2 different feature extractors (ResNet-50 pretrained with supervised learning on ImageNet cite ResNet or with SSL on TCGA) and 3 multiple instance learning models (MeanPool, Chowder and DeepMIL).

ImageNet feature extraction We first extracted features using the last layer of a ResNet-50 pretrained using supervised learning on the ImageNet-1k dataset. We used an auto-encoder to reduce dimension to $D = 256$ in order to overcome stability issues.

MoCo feature extraction Following Dehaene et al. (2020), we trained several ResNet-50 models using Momentum Contrast v2 (MoCo v2[Chen et al. (2020c)]). We used the exact same parameters and data augmentation scheme but a bigger ResNet backbone was used (the bottleneck number of channels is twice larger in every block). Three different feature extractors were trained: MoCo-CRC using 4.7M tiles from TCGA-CRC, MoCo-Gastric

	Size	MSI positive	Location	Origin
TCGA-CRC	555	78 (14%)	Colon (74%) - Rectum (26%)	US, 36 centers
TCGA-CRC-Kather	360	65 (18%)	Colon (74%)- Rectum (26%)	US, 34 centers
TCGA-Gastric	375	64 (17%)	Gastric	US, 22 centers
TCGA-Gastric-Kather	284	60 (21%)	Gastric	US, 20 centers
PAIP	47	12 (26 %)	Colon	Korea, 3 centers

Table 1: Datasets used in this study.

using 4.7M tiles from TCGA-Gastric, and MoCo-CRC-Gastric using the concatenation of the two previous datasets. Each model was trained for 200 epochs (approximately 30 hours) on 16 NVIDIA Tesla V100. The obtained MoCo features have a dimension $D = 2048$.

MeanPool As a baseline multiple instance learning method, we used a simple average pooling of the tile features followed by a logistic regression either with or without $L2$ penalization ($C = 0$, 0.5 or 1).

Chowder We implemented a variant of Chowder [Courtiol et al. (2018)]. A multilayer perceptron (MLP) with 128 hidden neurons and sigmoid activation is applied to each tile's features to output one score. The R ($R = 10$, 25 or 100) top and bottom scores are then concatenated and fed into a MLP with 128 and 64 hidden neurons and sigmoid activations.

DeepMIL We reimplemented the attention based model proposed by Ilse et al. (2018). A linear layer with N neurons ($N = 64$, 128 or 256 here) is applied to the embedding followed by a Gated Attention layer with N hidden neurons. A linear layer followed by a sigmoid activation is then applied to the output.

For both Chowder and DeepMIL, a random subset of $n = 8000$ tiles per WSI was used to accelerate training. All hyperparameters were tuned on the different training sets (see Results). Chowder and DeepMIL were trained with a learning rate of 0.001 using cross-entropy loss and the Adam optimizer [Kingma and Ba (2014)]. When multiple WSI were available for a given patient, we averaged the predictions of the models at test time.

3.2 Datasets

Three different cohorts were used in this study and are summarized in Table 1. For all cohorts, only FFPE images were used.

TCGA-CRC is a dataset of $n = 555$ patients from 36 centers in the US with colorectal tumors. It is a combination of two cohorts from TCGA: TCGA-COAD (colon adenocarcinoma) and TCGA-READ (rectum adenocarcinoma). TCGA-STAD (stomach adenocarcinoma), later referred as TCGA-Gastric, is a dataset of $n = 375$ patients from 22 centers in the US with gastric cancer. For both datasets, MSS/MSI-H labels defined by PCR assays were retrieved using TCGA-biolinks [Colaprico et al. (2016)]. As recommended by ESMO guidelines [Luchini et al. (2019)], MSI-L patients were classified as MSS. TCGA-CRC-Kather and TCGA-Gastric-Kather are variants of respectively TCGA-CRC and TCGA-Gastric datasets published by Kather et al. (2019). They were used here for

Methods	TCGA-CRC-Kather	TCGA-Gastric-Kather
Kather et al. (2019)	0.77 (0.62-0.87)	0.81 (0.69 - 0.90)
Yamashita et al. (2021)	0.82 (0.71-0.91)	-
Bilal et al. (2021)	0.90	-
Ours - MeanPool	0.85 (0.76 - 0.94)	0.78 (0.68 - 0.88)
Ours - Chowder	**0.92 (0.84 - 0.99)**	**0.83 (0.75 - 0.92)**
Ours - DeepMIL	0.85 (0.75 - 0.94)	0.79 (0.69 - 0.89)

Table 2: AUCs on Kather et al. (2019) train/test split. 95% CI are computed following [DeLong et al. (1988)] for our models, and using boostrapping for Kather et al. (2019) and Yamashita et al. (2021).

comparison purposes. These datasets consist of a lower number of cases because MSI-L patients were excluded. The exact same MSI labels were used.

PAIP (cohort from the Pathology AI Platform, http://www.wisepaip.org) is a dataset of $n = 47$ patients from 3 centers in Korea with colorectal tumors. MSS/MSI-H labels were determined using PCR assays.

4. Results

4.1 Cross-validations on TCGA-CRC and TCGA-Gastric

We first compared three multiple instance learning models, MeanPool, Chowder and DeepMIL, using the MoCo-CRC features on the TCGA-CRC-Kather cohort, and the MoCo-Gastric features on the TCGA-Gastric-Kather cohort. For the sake of comparison, we used the exact same train / test split as in [Kather et al. (2019)] and compared our results with the ones published in the literature on this split.

We tuned the hyperparameters of Chowder ($R = 10$, 25 or 100), DeepMIL (size of attention layer = 64, 128 or 256), MeanPool (l2 penalization = 0, 0.5 or 1) and the number of epochs (5, 10, 20, or 30) using a grid search on the training sets. For both TCGA-CRC-Kather and TCGA-Gastric-Kather, Chowder model performed best and obtained respectively AUCs of 0.92 and 0.83, achieving state-of-the-art results on these datasets (see Table 2).

To analyze the gain of MoCo features over ImageNet ones, we ran larger cross-validation experiments using 15 distinct splits on the full TCGA cohorts (5 fold cross-validation, repeated 3 times) and reported results in Table 3. For a fair comparison, we tuned the hyperparameters of all models on the training set of the Kather split. In all our experiments, MoCo substantially outperformed ImageNet in both TCGA-CRC and TCGA-Gastric cohorts. Results obtained with MoCo features were also better than the ones previously reported with cross-validation. We also report cross-validation results using center split in Supplementary Table S1.

	TCGA-CRC		TCGA-Gastric	
Echle et al. (2020)	0.74 (0.66–0.80)		-	
Kather et al. (2020)	-		0.72	
Bilal et al. (2021)	0.86		-	
	ImageNet	MoCo-CRC	ImageNet	MoCo-Gastric
Ours - MeanPool	0.84 (0.05)	0.87 (0.05) +0.03	0.76 (0.04)	0.82 (0.05) +0.06
Ours - Chowder	0.82 (0.06)	0.88 (0.04) +0.06	0.73 (0.07)	0.83 (0.06) +0.11
Ours - DeepMIL	0.82 (0.05)	0.88 (0.05) +0.06	0.74 (0.01)	0.85 (0.05) +0.11

Table 3: Performances for 3×5 folds cross-validation (AUC), means, and standard deviations on TCGA-CRC and TCGA-Gastric datasets. Mean, lower and upper bounds on 3 folds are reported by Echle et al. (2020), means on respectively 4 folds and 3 folds are reported by Bilal et al. (2021) and Kather et al. (2020)

Methods	ImageNet	MoCo-CRC
MeanPool	0.61 (0.42 - 0.80)	0.82 (0.66 - 0.99) +0.21
Chowder	0.82 (0.67 - 0.97)	0.97 (0.93 - 1.00) +0.15
DeepMIL	0.83 (0.67 - 1.0)	0.90 (0.78 - 1.00) +0.07

Table 4: AUCs on PAIP for models trained on TCGA-CRC. Bilal et al. (2021) report AUC of 0.98 but no CI interval is given.

4.2 External validation on CRC dataset PAIP

A limitation of the previous experiments is that ResNet using SSL were pretrained on the full TCGA cohorts, slightly breaking the train/test split independence assumption even if no MSI/MSS labels were used. Thus, we further evaluated MSI detection on an independent cohort of $n = 47$ colon cases from PAIP organisation. We used the median prediction of the ensemble of models trained during cross-validation in the previous experiment. Chowder with MoCo-CRC features yielded an AUC of 0.97 on par with the 0.98 reported in Bilal et al. (2021), and significantly higher than Chowder with ImageNet features (AUC of 0.82, $p = 0.03$ with DeLong's test [DeLong et al. (1988)]). The superiority of MoCo-CRC features was observed for all models (Table 4), confirming that the MoCo-CRC pretrained backbone is a more robust feature extractor than the ImageNet for histology images.

4.3 Transfer CRC to Gastric

We assessed the performances for MSI detection when transferring models trained on TCGA-CRC to TCGA-Gastric with three different feature extractors: ImageNet, MoCo-CRC, MoCo-CRC-Gastric. For fair comparison, all hyperparameters were tuned on the training split of TCGA-CRC-Kather (Supplementary Table S2).

MoCo-CRC-Gastric models consistently yielded the best results on TCGA-Gastric with AUCs up to 0.80 (Table 5) while also demonstrating high performances on TCGA-CRC

Methods	ImageNet	MoCo-CRC	MoCo-CRC-Gastric
MeanPool	0.76 (0.69-0.82)	0.67 (0.59-0.74)	0.76 (0.69-0.82)
Chowder	0.72 (0.66-0.78)	0.73 (0.66-0.80)	0.78 (0.73-0.84)
DeepMIL	0.66 (0.58-0.73)	0.71 (0.64-0.79)	0.80 (0.75-0.86)

Table 5: AUCs for models trained on TCGA-CRC and evaluated on TCGA-Gastric with different feature extractors.

(Supplementary Table S3). Surprisingly, the ImageNet MeanPool model, which performed poorly on PAIP, also obtained a high AUC of 0.76, however significantly lower compared to 0.80 ($p = 0.05$ with DeLong's test). To our knowledge, this is the best performance reported in a transfer setting from one organ to another.

5. Interpretability

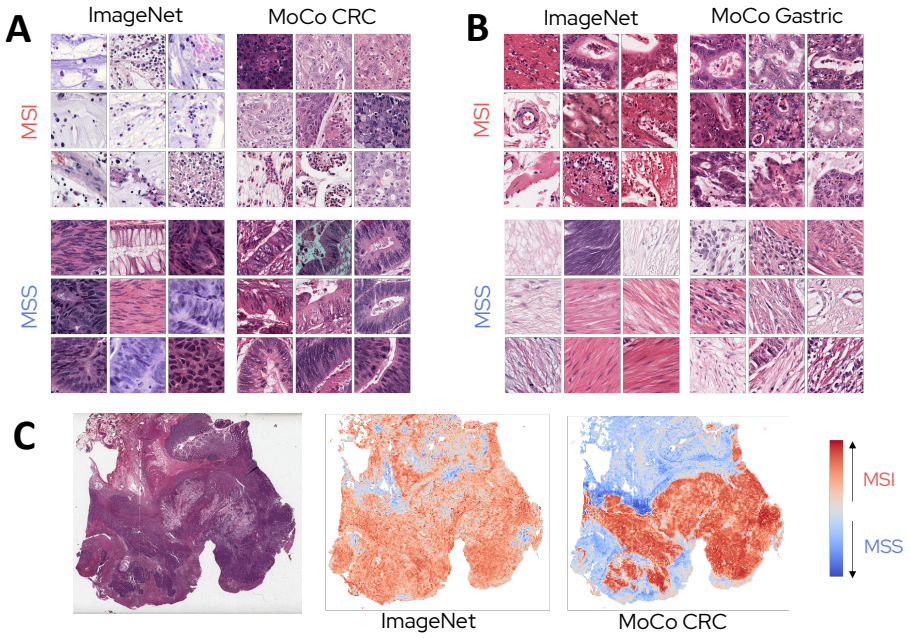

Figure 2: Most predictive regions identified by Chowder A. Tiles most predictive of MSI and MSS phenotypes in CRC. B. Similar for gastric cancer. C. Heatmaps obtained with ImageNet or MoCo-CRC features.

Within Chowder, each tile is associated with a single score and a MLP is applied to the top and lowest scores of the WSI (Section 3.1). To explore potential differences in interpretability when using ImageNet or MoCo features, a pathologist expert in dMMR tumors reviewed the subset of tiles associated with extreme scores within TCGA-CRC and TCGA-Gastric cohorts, for both feature extractors.

Among the top scored tiles (Figure 2 A, B), the pathologist recognized histological patterns which have been previously described as associated with MSI tumors, including tumor lymphocyte infiltration, mucinous differentiation and presence of dirty necrosis - the latter being more frequently detected in gastric tumors. While ImageNet-based models outputs were based on patterns within and outside tumor regions (such as lymphocyte infiltration in the tumor microenvironment), MoCo-based models clearly focused on tumor epithelium, highlighting both poor differentiation of epithelial cells and tumor infiltrating lymphocytes (TILs) as patterns associated with MSI tumors. Specifically, MoCo models better segmented tumoral regions, as shown by the tile scores heatmap at the WSI level (Figure 2 C).

Overall, lowest scored tiles were unrelated to known MSI patterns, displaying non tumoral tissues such as muscle cells. In the TCGA-CRC cohort, the MoCo-CRC model additionally highlighted the presence of intratumoral epithelial cells without MSI tumor characteristics, showing once again a higher focus on tumor regions, in line with the regions actually inspected by a pathologist when searching for MSI related patterns.

6. Discussion

In this work, we showed that feature extractors pretrained using SSL (MoCo V2) on TCGA reach state-of-the-art results for MSI prediction both in colorectal and gastric cancers (Table 2). Extensive cross-validations showed the clear superiority of these models over their counterparts pretrained using ImageNet (Table 3). In addition, the former generalize better on an external CRC cohort (Table 4). Finally, we observed that using a feature extractor pretrained on several organs using SSL (Table 5) opens the way to both state-of-the-art performances in cross-validation and robust generalization from one organ to another.

There are several limitations to our current study. First, our models could benefit from being trained on more data as shown in the training curve from Echle et al. (2020) (Figure 1.c). Second, our models should be validated on larger cohorts, encompassing different patient populations, treatments (neoadjuvant chemotherapy can impact cell morphology), scanner manufacturer and sample types (resections, biopsies). Third, SSL techniques are evolving rapidly using new architectures such as Vision Transformers [Dosovitskiy et al. (2020)] and more experiments are required to find how to apply them in histology, including taking into account the spatial arrangement of the tiles. However, such experiments require access to extensive computational resources, which limits reproducibility. Finally, in contrast to several concurrent works [Echle et al. (2020); Bilal et al. (2021); Kather et al. (2019)] that fine-tuned the backbones, our study kept them entirely frozen in this study.

A recent study, Kacew et al. (2021) showed that performances of deep learning models may impact the diagnosis of MSI for patients with CRC. Notably, MSI diagnosis is not routinely done for patients with other solid tumors, missing the identification of candidate patients for immunotherapy. Our results indicate that SSL is a promising solution to develop accurate models for frequent tumor localization such as oesophagus, pancreas, small intestine or even brain, where images are available but MSI prevalence is too low for systematic IHC or molecular testing.

Acknowledgments

The results published here are part based upon data generated by the TCGA Research Network: https://www.cancer.gov/tcga.

Regarding the PAIP dataset: 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).

Supplementary Tables

		TCGA-CRC		TCGA-Gastric	
	Split	ImageNet	MoCo-CRC	ImageNet	MoCo-Gastric
MeanPool	CV	0.84 (0.05)	0.87 (0.05) +0.03	0.76 (0.04)	0.82 (0.05) +0.06
MeanPool	CV centers	0.78 (0.10)	0.85 (0.07) +0.07	0.72 (0.12)	0.85 (0.07) +0.13
DeepMIL	CV	0.82 (0.05)	0.88 (0.05) +0.06	0.74 (0.01)	0.85 (0.05) +0.11
DeepMIL	CV centers	0.79 (0.059)	0.84 (0.11) +0.05	0.73 (0.15)	0.85 (0.05) +0.12
Chowder	CV	0.82 (0.06)	0.88 (0.04) +0.06	0.73 (0.07)	0.84 (0.06) +0.11
Chowder	CV centers	0.73 (0.14)	0.83 (0.12) +0.10	0.72 (0.08)	0.86 (0.05) +0.14

Table S1: Cross-validation performances (AUC) on TCGA-CRC and TCGA-Gastric. We report mean and standard deviation on 3×5 folds. We split the data into 5 fold either randomly (CV) or by ensuring that all samples from a center are in the same set (CV centers).

Feature extractor	Train dataset	MeanPool	Chowder	DeepMIL
ImageNet	CRC	C = 1	R = 10 30 epochs	N = 64 30 epochs
ImageNet	Gastric	C = 0.0	R = 25 30 epochs	N = 64 20 epochs
MoCo-CRC	CRC	C = 0.5	R = 10 10 epochs	N = 32 10 epochs
MoCo-Gastric	Gastric	C = 1.0	R = 100 30 epochs	N = 128 10 epochs
MoCo-CRC-Gastric	CRC	C = 0.5	R = 100 30 epochs	N = 64 10 epochs

Table S2: Hyperparameters used for the different models. All hyperparameters were tuned on the training sets of the TCGA-CRC-Kather or TCGA-Gastric-Kather cohorts. C refers to L2 penalization coefficient, R to the number of extreme tiles used in Chowder and N to the size of the attention layer in DeepMIL.

Methods	MoCo-CRC-Gastric
MeanPool	0.86 (0.04)
Chowder	0.88 (0.05)
DeepMIL	0.88 (0.06)

Table S3: Cross-validation performances (AUC) on TCGA-CRC, for the models trained using the features of MoCo-CRC-Gastric. The result reported are the mean and the standard deviation from 5-fold cross-validation repeated 3 times.

References

Christian Abbet, Linda Studer, Andreas Fischer, Heather Dawson, Inti Zlobec, Behzad Bozorgtabar, and Jean-Philippe Thiran. Self-rule to adapt: Learning generalized features from sparsely-labeled data using unsupervised domain adaptation for colorectal cancer tissue phenotyping. In *Medical Imaging with Deep Learning*, 2021.

Mohsin Bilal, Shan E Ahmed Raza, Ayesha Azam, Simon Graham, Muhammad Ilyas, Ian A Cree, David Snead, Fayyaz Minhas, and Nasir M Rajpoot. Novel deep learning algorithm predicts the status of molecular pathways and key mutations in colorectal cancer from routine histology images. *medRxiv*, 2021.

Rui Cao, Fan Yang, Si-Cong Ma, Li Liu, Yu Zhao, Yan Li, De-Hua Wu, Tongxin Wang, Wei-Jia Lu, Wei-Jing Cai, et al. Development and interpretation of a pathomics-based model for the prediction of microsatellite instability in colorectal cancer. *Theranostics*, 10(24):11080, 2020.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. *arXiv preprint arXiv:2104.14294*, 2021.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020a.

Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey Hinton. Big self-supervised models are strong semi-supervised learners. *arXiv preprint arXiv:2006.10029*, 2020b.

Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020c.

Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. *arXiv preprint arXiv:2104.02057*, 2021.

Antonio Colaprico, Tiago C Silva, Catharina Olsen, Luciano Garofano, Claudia Cava, Davide Garolini, Thais S Sabedot, Tathiane M Malta, Stefano M Pagnotta, Isabella Castiglioni, et al. Tcgabiolinks: an r/bioconductor package for integrative analysis of tcga data. *Nucleic acids research*, 44(8):e71–e71, 2016.

Pierre Courtiol, Eric W Tramel, Marc Sanselme, and Gilles Wainrib. Classification and disease localization in histopathology using only global labels: A weakly-supervised approach. *arXiv preprint arXiv:1802.02212*, 2018.

Olivier Dehaene, Axel Camara, Olivier Moindrot, Axel de Lavergne, and Pierre Courtiol. Self-supervision closes the gap between weak and strong supervision in histology. *arXiv preprint arXiv:2012.03583*, 2020.

Elizabeth R DeLong, David M DeLong, and Daniel L Clarke-Pearson. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*, pages 837–845, 1988.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Amelie Echle, Heike Irmgard Grabsch, Philip Quirke, Piet A van den Brandt, Nicholas P West, Gordon GA Hutchins, Lara R Heij, Xiuxiang Tan, Susan D Richman, Jeremias Krause, et al. Clinical-grade detection of microsatellite instability in colorectal tumors by deep learning. *Gastroenterology*, 159(4):1406–1416, 2020.

Kenji Fujiyoshi, Tatsuro Yamaguchi, Miho Kakuta, Akemi Takahashi, Yoshiko Arai, Mina Yamada, Gou Yamamoto, Sachiko Ohde, Misato Takao, Shin-ichiro Horiguchi, et al. Predictive model for high-frequency microsatellite instability in colorectal cancer patients over 50 years of age. *Cancer medicine*, 6(6):1255–1263, 2017.

Jacob Gildenblat and Eldad Klaiman. Self-supervised similarity learning for digital pathology. *arXiv preprint arXiv:1905.08139*, 2019.

Joel K Greenson, Shu-Chen Huang, Casey Herron, Victor Moreno, Joseph D Bonner, Lynn P Tomsho, Ofer Ben-Izhak, Hector I Cohen, Phillip Trougouboff, Jacob Bejhar, et al. Pathologic predictors of microsatellite instability in colorectal cancer. *The American journal of surgical pathology*, 33(1):126, 2009.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent: A new approach to self-supervised learning. *arXiv preprint arXiv:2006.07733*, 2020.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738, 2020.

Lindsey A Hildebrand, Colin J Pierce, Michael Dennis, Munizay Paracha, and Asaf Maoz. Artificial intelligence for histology-based detection of microsatellite instability and prediction of response to immunotherapy in colorectal cancer. *Cancers*, 13(3):391, 2021.

Runyu Hong, Wenke Liu, Deborah DeLair, Narges Razavian, and David Fenyö. Predicting endometrial cancer subtypes and molecular features from histopathology images using multi-resolution deep learning models. *bioRxiv*, 2020.

Angela Hyde, Daniel Fontaine, Susan Stuckless, Roger Green, Aaron Pollett, Michelle Simms, Payal Sipahimalani, Patrick Parfrey, and Banfield Younghusband. A histology-based model for predicting microsatellite instability in colorectal cancers. *The American journal of surgical pathology*, 34(12):1820–1829, 2010.

Maximilian Ilse, Jakub Tomczak, and Max Welling. Attention-based deep multiple instance learning. In *International conference on machine learning*, pages 2127–2136. PMLR, 2018.

Mark A Jenkins, Shinichi Hayashi, Anne-Marie O'shea, Lawrence J Burgart, Tom C Smyrk, David Shimizu, Paul M Waring, Andrew R Ruszkiewicz, Aaron F Pollett, Mark Redston, et al. Pathology features in bethesda guidelines predict colorectal cancer microsatellite instability: a population-based study. *Gastroenterology*, 133(1):48–56, 2007.

Alec J Kacew, Garth W Strohbehn, Loren Saulsberry, Neda Laiteerapong, Nicole A Cipriani, Jakob N Kather, and Alexander T Pearson. Artificial intelligence can cut costs while maintaining accuracy in colorectal cancer genotyping. *Frontiers in Oncology*, 11, 2021.

Jakob Nikolas Kather, Alexander T Pearson, Niels Halama, Dirk Jäger, Jeremias Krause, Sven H Loosen, Alexander Marx, Peter Boor, Frank Tacke, Ulf Peter Neumann, et al. Deep learning can predict microsatellite instability directly from histology in gastrointestinal cancer. *Nature medicine*, 25(7):1054–1056, 2019.

Jakob Nikolas Kather, Lara R Heij, Heike I Grabsch, Chiara Loeffler, Amelie Echle, Hannah Sophie Muti, Jeremias Krause, Jan M Niehues, Kai AJ Sommer, Peter Bankhead, et al. Pan-cancer image-based detection of clinically actionable genetic alterations. *Nature Cancer*, 1(8):789–799, 2020.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Navid Alemi Koohbanani, Balagopal Unnikrishnan, Syed Ali Khurram, Pavitra Krishnaswamy, and Nasir Rajpoot. Self-path: Self-supervision for classification of pathology images with limited annotations. *IEEE Transactions on Medical Imaging*, 2021.

Dung T Le, Jennifer N Durham, Kellie N Smith, Hao Wang, Bjarne R Bartlett, Laveet K Aulakh, Steve Lu, Holly Kemberling, Cara Wilt, Brandon S Luber, et al. Mismatch repair deficiency predicts response of solid tumors to pd-1 blockade. *Science*, 357(6349): 409–413, 2017.

Sung Hak Lee, In Hye Song, and Hyun-Jong Jang. Feasibility of deep learning-based fully automated classification of microsatellite instability in tissue slides of colorectal cancer. *International Journal of Cancer*, 2021.

Bin Li, Yin Li, and Kevin W Eliceiri. Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14318–14328, 2021a.

Chunyuan Li, Jianwei Yang, Pengchuan Zhang, Mei Gao, Bin Xiao, Xiyang Dai, Lu Yuan, and Jianfeng Gao. Efficient self-supervised vision transformers for representation learning. *arXiv preprint arXiv:2106.09785*, 2021b.

Ming Y Lu, Richard J Chen, Jingwen Wang, Debora Dillon, and Faisal Mahmood. Semi-supervised histology classification using deep multiple instance learning and contrastive predictive coding. *arXiv preprint arXiv:1910.10825*, 2019.

C Luchini, F Bibeau, MJL Ligtenberg, N Singh, A Nottegar, T Bosse, R Miller, N Riaz, J-Y Douillard, F Andre, 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. *Annals of Oncology*, 30(8): 1232–1243, 2019.

Vinay Prasad, Victoria Kaestner, and Sham Mailankody. Cancer drugs approved based on biomarkers and not tumor type—fda approval of pembrolizumab for mismatch repair-deficient solid cancers. *JAMA oncology*, 4(2):157–158, 2018.

Ruth Román, Montse Verdú, Miquel Calvo, August Vidal, Xavier Sanjuan, Mireya Jimeno, Antonio Salas, Josefina Autonell, Isabel Trias, Marta González, et al. Microsatellite instability of the colorectal carcinoma can be predicted in the conventional pathologic examination. a prospective multicentric study and the statistical analysis of 615 cases consolidate our previously proposed logistic regression model. *Virchows Archiv*, 456(5): 533–541, 2010.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

Daniel J Sargent, Silvia Marsoni, Genevieve Monges, Stephen N Thibodeau, Roberto Labianca, Stanley R Hamilton, Amy J French, Brian Kabat, Nathan R Foster, Valter Torri, et al. Defective mismatch repair as a predictive marker for lack of efficacy of fluorouracil-based adjuvant therapy in colon cancer. *Journal of Clinical Oncology*, 28(20):3219, 2010.

Chetan L Srinidhi, Seung Wook Kim, Fu-Der Chen, and Anne L Martel. Self-supervised driven consistency training for annotation efficient histopathology image analysis. *arXiv preprint arXiv:2102.03897*, 2021.

Magali Svrcek, Olivier Lascols, Romain Cohen, Ada Collura, Vincent Jonchère, Jean-François Fléjou, Olivier Buhard, and Alex Duval. Msi/mmr-deficient tumor diagnosis: Which standard for screening and for diagnosis? diagnostic modalities for the colon and other sites: Differences between tumors. *Bulletin du cancer*, 106(2):119–128, 2019.

Rikiya Yamashita, Jin Long, Teri Longacre, Lan Peng, Gerald Berry, Brock Martin, John Higgins, Daniel L Rubin, and Jeanne Shen. Deep learning model for the prediction of microsatellite instability in colorectal cancer: a diagnostic study. *The Lancet Oncology*, 22(1):132–141, 2021.

Renyu Zhang, Boleslaw L Osinski, Timothy J Taxter, Jason Perera, Denise J Lau, and Aly A Khan. Adversarial deep learning for microsatellite instability prediction from histopathology slides. In *Proceedings of the 1st Conference on Medical Imaging with Deep Learning (MIDL 2018), Amsterdam, The Netherlands*, pages 4–6, 2018.