Predict Microsatellite Instability Status from Histological Images of Colorectal Cancer using Deep Learning

Ruoxun Zi and **Shengjia Chen**
Department of Research and Science,
New York University School of Medicine,
New York, NY 10016-3240, USA

Abstract

Microsatellite instability (MSI) is an effective biomarker to determine the therapy for patients with colorectal cancer. However, current MSI identification tests are not available for all patients. To overcome this problem, we developed a deep learning based pipeline, combining convolutional neural network, multiple instance learning and recurrent neural network, to predict the microsatellite status directly from histopathology images. The methods have been tested on the Cancer Genome Atlas cohort with an accuracy of 0.86 and an AUROC score of 0.90. Our results show that the model can effectively predict MSI from histopathology images and shows the broad application of deep learning to aid clinical diagnosis. Code available: MSI_vs_MSS_Classification

Keywords: Microsatellite instability, Convolutional neural network, Multiple instance learning

1. Introduction

Colorectal cancer (CRC) is the third most common and second deadly cancer worldwide. Microsatellite instability is a molecular maker of deficient DNA mismatch repair that is found in approximately 15% of CRC patients (Vilar and Gruber, 2010). MSI has diagnostic, prognostic and therapeutic implications in CRC and other cancers. Testing all CRC patients for MSI is recommended by multiple professional societies, which requires either an immunohistochemical analysis or a PCR-based assay (Hildebrand et al., 2021). However, universal testing for MSI has not been implemented due to cost and resource limitations, which raises an unmet need for a more affordable and accessible way to detect MSI status.

Deep learning is being explored for a myriad of research and clinical use, with broad applications to digital pathology in oncology. Recently, several studies have investigated the potential of deep learning to predict MSI directly from hematoxylin and eosin (H&E) stained slides. Typically, a two-step approach was used, consisting of a tile-level prediction and a slide-level diagnosis. ResNet-18 was selected in many studies to assign a MSI probability on the tile-level (Kather et al., 2019; Cao et al., 2020). Different methods were used to classify MSI in the whole-slide image (WSI) level. Kather et al. determined the MSI status of each slide by the majority of its constituent tiles (Kather et al., 2019). Cao et al. used machine learning methods to aggregate the tile probabilities for WSI classification, consisting of a patch likelihood histogram pipeline and bag of words pipeline, which reported an improved overall predictive accuracy (Cao et al., 2020).

The main hypothesis is that deep learning can be used to predict the MSI from the WSI. To be specific, the testable hypothesis is that the performance of the models used in this work is comparable to other reported works. In this work, we designed a two-step classification model, consisting of convolutional neural network (CNN) and recurrent neural network (RNN), to predict MSI versus microsatellite stability (MSS) from WSI.

2. Methods

2.1 Data

The dataset used in this study contains 192,312 image tiles cut from WSIs of 374 CRC patients in the cohort of The Cancer Genome Atlas (TCGA). The data preprocessing, including tumor detection, tile cutting and color normalization, has been done by other researchers (Kather, 2019). The label was assigned for each WSI, which is MSI or MSS. The dataset was split into training, validation and testing by 70%, 15% and 15% on a slide-level. The training set is balanced with the same number of MSI and MSS tiles. However, the dataset is imbalanced on the slide-level, with the ratio of 3:1 for MSS vs. MSI (Extended Data Table 1). Each tile is in the size of 224×224×3 pixels as an RGB image.

2.2 Model

2.2.1 Pipeline

The pipeline of the proposed model is shown in Fig 1. In the first step, a CNN model was implemented for the tile-level classification, where the input is a tile, and the output is the probability of being MSI for each tile. Widely used CNN models for medical imaging classification tasks were compared for this task. Multiple instance learning was implemented to compensate for the lack of tile-level labels. In the second step, a slide-level classification was done by aggregation of the information of all tiles from the same slide. Different aggregation methods were implemented, including majority vote, machine learning-based methods and RNN.

All experiments were conducted on Google Cloud Platform cluster. Each model was trained on a single GPU. PyTorch (torch1.10.2, torchvision0.11.3) with PyTorch-lightning (v1.6.0) was used for building and training model. All evaluation metrics in the training and validation process were automatically tracked with Tensorboard.

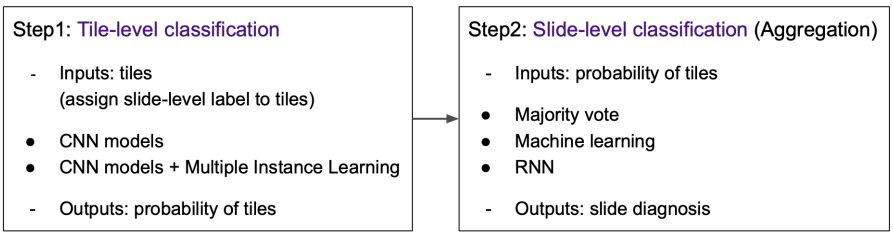


Figure 1: Pipeline of the proposed methods.

2.2.2 Tile-level Classification

For tile-level classification, six different CNN models were trained to find the best model for this work, including AlexNet, Resnet-18/34, VGG-11BN, InceptionV3, and DenseNet121. All CNN models were pretrained on the ImageNet and both CNN and classifier layers are trainable. Adam optimizer was used for training with a learning rate of 1e-3 and 1e-4, a batch size of 64 and epochs of 50. Cross-entropy was selected as the loss function. To avoid overfitting, the model giving the highest validation accuracy was saved as the best model during the training process.

Due to the lack of a true label for each tile, it is unfeasible to rely on the normal supervised training procedure (dash box in Fig 2a). In addition, we know that if the slide is MSS, all tiles must be MSS. On the contrary, if the slide is MSI, at least one of the possible tiles must be MSI. This formalization of the WSI classification problem is an example of the general standard multiple instance assumption, for which multiple instance learning (MIL) is a widely used solution. Therefore, MIL was integrated in the training procedure, which includes a full inference pass through the dataset, ranking the tiles according to their probability of being MSI, and learning on the top-ranking tiles per slide (Fig 2a).

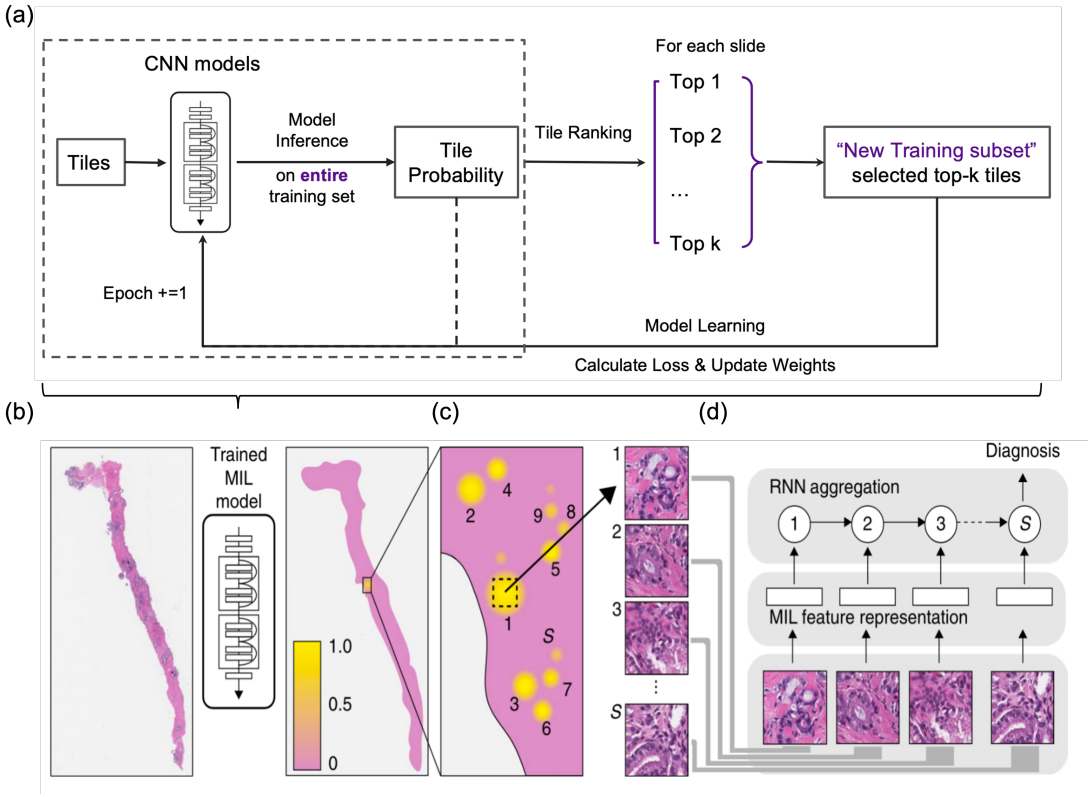


Figure 2: The pipeline of CNN modeling training under MIL followed by the RNN aggregation. (a) MIL training process, (b) tile-level prediction, (c) preparation of the RNN input, and (d) RNN aggregation process.

2.2.3 Slide-level Classification

Three different methods were implemented for the slide-level classification. Firstly, the predicted MSI status for each slide was determined by the predicted MSI status of the majority of its constituent tiles (Kather et al., 2019), which was used as the baseline for this work. Secondly, we tried a machine learning based method, consisting of a Naïve Bayes classifier on the probabilities of tiles from the same slide and a gradient boosting classifier on the histogram of the probabilities, followed by a weighted sum to ensemble the probabilities of two machine learning classifiers as the final prediction (Cao et al., 2020). The first two methods for aggregation were applied to all CNN models discussed in 2.2.2.

Lastly, a RNN was trained, which is suited for the CNN models training with MIL (Campanella et al., 2019). After training a tile-level CNN model with MIL (Fig. 2b), the probability for each tile was known, which was used to select the top-k most suspicious tiles for each slide (Fig. 2c). These selected top-k tiles were used as the input for the aggregation RNN, with the size of k×224×224×3. The k most suspicious tiles in each slide were sequentially passed to the RNN to predict the final slide-level classification (Fig 2c). In the RNN, an encoder was used to extract the features from each tile, The trained tile-level MIL model was selected as the encoder, which can utilize the features learned in the first step.

2.3 Evaluation Metrics

The accuracy and area under the receiver operating characteristic curve (AUROC) are used as the performance metrics to evaluate our binary classification model both on the tile-level and slide-level. Confusion matrix for the best model is also reported due to data imbalance.

3. Results

3.1 Model Training

In Extended Data Fig 1, the training loss decreased and got converged during the training process with a training accuracy of 0.99 at the last epoch, which indicates the effective learning process. The validation accuracy increased at the first several epochs and then started fluctuation.

3.2 Comparison of CNN Models

The best performance of each CNN model after fine-tuning the learning rate is shown in Table 1. Densenet outperforms other five models on both the tile-level and slide-level, with the best accuracy of 0.86 and an AUROC score of 0.90. Resnet18 with aggregation of majority vote gave an accuracy of 0.80 and an AUROC score of 0.78, which is slightly lower than the value reported in Kather(citation) using the same pipeline. Resnet18 with machine learning-based aggregation gave an accuracy of 0.86 and an AUROC score of 0.85, which is comparable to the value reported in Cao(citation) using the same pipeline. The ROC curves are smooth on the tile-level due to the large number of tiles and are in a stepwise shape on the slide-level as expected (Fig 3).

Table 1: Performance of different CNN models

Model	Tile Level		Slide Level			
			Baseline		ML	
	Accuracy	AUROC	Accuracy	AUROC	Accuracy	AUROC
Alex	0.69	0.63	0.72	0.78	0.74	0.76
Resnet18	0.72	0.67	0.8	0.78	**0.86**	**0.85**
Resnet34	0.72	0.69	0.78	0.77	0.8	0.8
VGG	0.76	0.74	0.78	0.77	0.8	0.82
Inception	0.74	0.71	0.76	0.75	0.74	0.81
Densenet	**0.75**	**0.8**	**0.84**	**0.87**	**0.86**	**0.9**

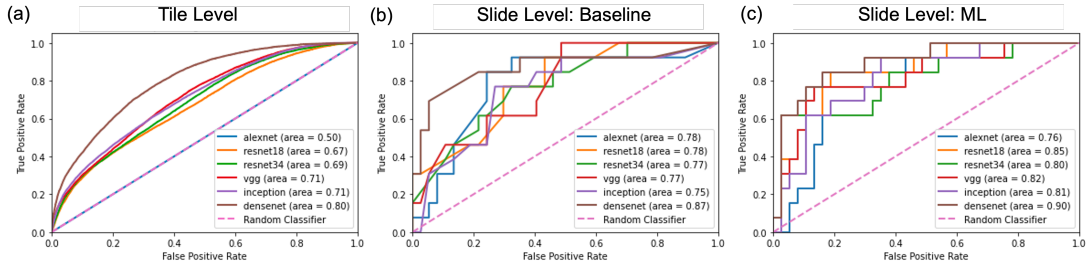


Figure 3: ROC curves of different CNN models at (a) tile-level, (b) followed by a majority vote aggregation, (c) followed by a machine learning based aggregation.

3.3 Comparison of MIL Training

The MIL and its following RNN aggregation were only tested on the selected two models: Resnet18 and Densenet. The results of Densenet are shown in Fig 4 and Table 2 and of Resnet18 are shown in Extended Data Table 2 and Fig 3 with similar results. For each aggregation method, using MIL training provided a lower accuracy and AUROC. For the same model, whether training with or without MIL, RNN aggregation provided a lower accuracy and AUROC than the other two aggregation methods. The Confusion matrix of the best model (Densenet + machine learning aggreagation) is shown in Fig 4b.

Table 2: Performance of Densenet training with and without MIL

Metrics	Model	Tile Level	Slide Level		
			Baseline	ML	RNN
Accuracy	Densenet	0.75	0.84	**0.86**	0.64
	Densenet + MIL	0.73	0.74	0.74	0.72
AUROC	Densenet	0.8	0.87	**0.9**	0.75
	Densenet + MIL	0.63	0.74	0.8	0.64

4. Discussion

In this work, we developed a deep learning pipeline that combines CNN under a MIL training approach with three different aggregation approaches, including majority vote, machine learning and RNN. The best model is Densenet combining machine learning aggregation, which obtained an accuracy of 0.86 and an AUROC score of 0.90. The performance is comparable to the values in other papers, which confirmed our hypothesis. Resnet18, as a popular model for tile-level classification used in other papers, gave the second highest accuracy and AUROC. Furthermore, Densenet performed better than Resnet18. To our knowledge, there is no work reported using Densenet for the prediction of MSI status.

Training under a MIL approach didn't help improve the performance of the model as expected. The original idea of MIL was applied for tumor detection (Campanella et al., 2019), which was proposed to solve the problem of lack of tile-level labels. Since the same application on pathological images and the same multiple instance assumption, it is possible to transfer to the current task for MSI status detection. A tradeoff between making the

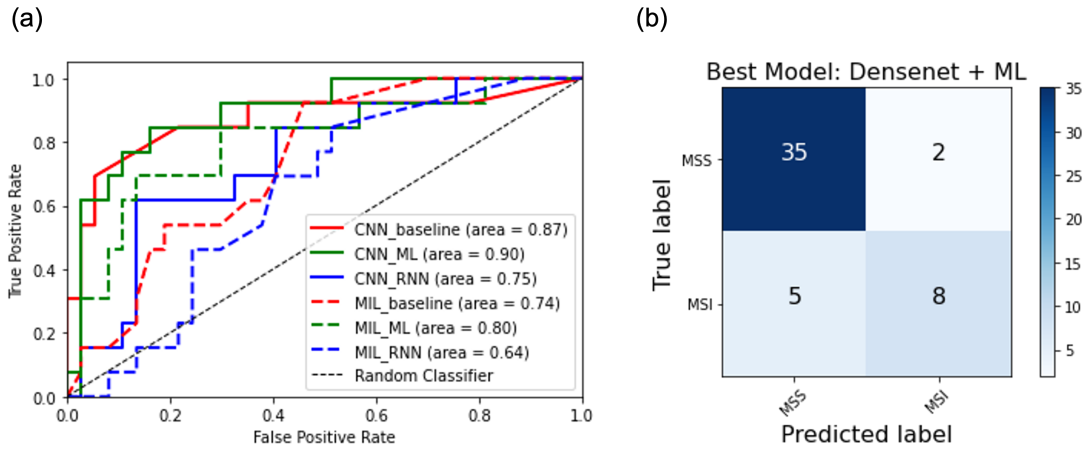


Figure 4: Performance of Densenet: (a) ROC curves with and without MIL training, and (b) confusion matrix of best model.

network more sensitive to detect MSI and training the network using a subset of the training set occurs when training under MIL. However, training by a subset is especially problematic for this work due to the limited number of slides with imbalanced labels, which is also the reason for the failure of RNN. On the other hand, it is observed that on the tile level the rate of true positive was increased when applying MIL in the training (Extended Data Fig 3c,d). It implies that it is possible to improve the performance using a dataset containing a larger number of slides with a balanced microsatellite status for the training under the MIL approach and followed by the RNN aggregation. Furthermore, hyperparameter tuning, such as the choice of k and the choice of the weight of the loss function could help mitigate the problem of imbalanced data. Currently, the maximum value of k was chosen as 5 without removing any slides.

All pretrained models were obtained from ImageNet. The CNN model trained without MIL was used as another pretrained model of the model training under MIL, which shows an improved performance compared to the one using ImageNet. Therefore, it is possible to improve the performance of the network further using other models trained on pathological images as the pretrained model.

Training on a large and balanced dataset will be the next step to improve the performance. Testing on a dataset from other cohorts or in different types of cancer would be a good way to evaluate the robustness of this pipeline. The tile cutting has already been done for the dataset we used in this work, in which the spatial information of the tile within the slides is lost. As a future work, cutting WSI into tiles at our end will give us this kind of information, which will enable the visualization (e.g. predicted tile-level MSI map of the slide) and localization of the regions with MSI status. Furthermore, integrating the spatial information into the aggregation approaches could help the slide-level diagnosis. To this end, an end-to-end pipeline will be developed from WSIs to diagnosis.

Data availability

Training images for MSI detection are obtained from the published paper: Kather, Jakob Nikolas, et al. "Deep learning can predict microsatellite instability directly from histology in gastrointestinal cancer." Nature medicine 25.7 (2019): 1054-1056. They are available at `https://doi.org/10.5281/zenodo.2530835`.

Code availability

The source code of this work can be downloaded from github:
`https://github.com/Gaskell-1206/MSI_vs_MSS_Classification`.
All training procedures are shown in tensorboard:
`https://tensorboard.dev/experiment/ZHvEWuBmT1KwO8NbEORHMg/#scalars`.

Contributions

Shengjia Chen built and trained the CNN models and CNN models training under a MIL approach. Ruoxun Zi helped to clarify the workflow of MIL and helped to debug. Ruoxun Zi wrote the codes of the three methods of aggregation (majority vote, machine learning and RNN). Shengjia helped to debug the RNN. Presentation and final report were completed together.

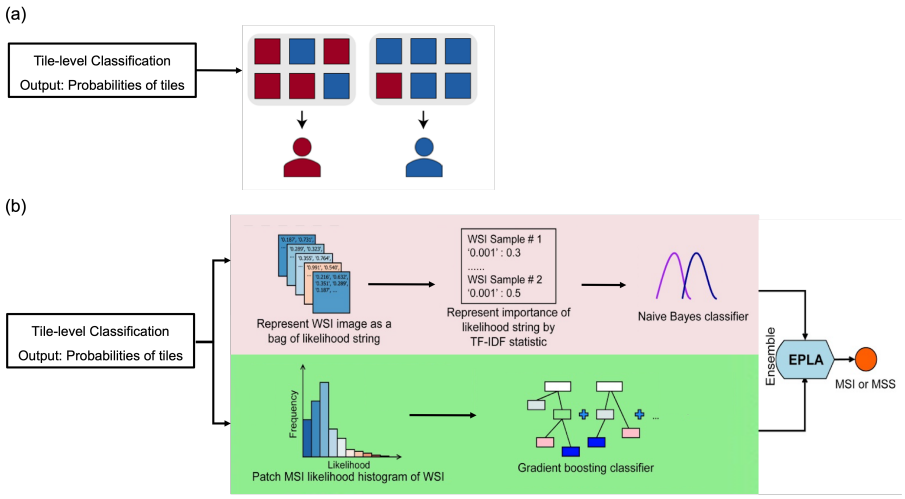
Appendix

Extended Table 1: Summary of dataset

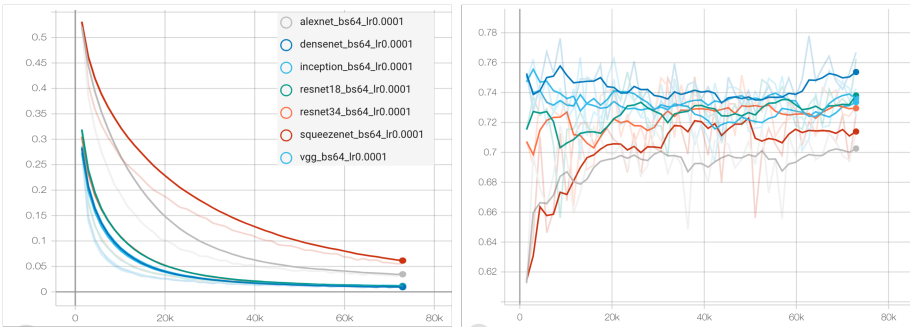
Data	SET	Label	Slides	Slides in Total	Tiles	Tiles in Total
CRC	Train	MSI	40	273	46,704	93,408
	Train	MSS	233		46,704	
	Val	MSI	13	51	15,489	51,801
	Val	MSS	38		36,312	
	Test	MSI	13	50	12,846	47,103
	Test	MSS	37		34,257	

Extended Table 2: Performance of Resnet training with and without MIL

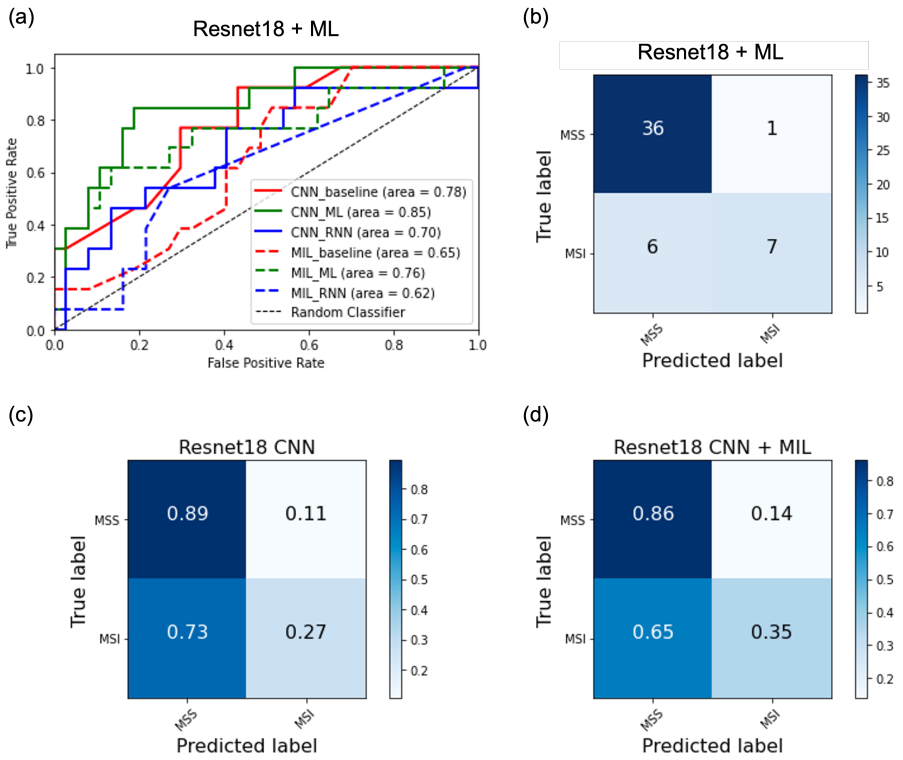
Metrics	Model	Tile Level	Slide Level		
			Baseline	ML	RNN
Accuracy	Resnet18	0.72	0.8	**0.86**	0.56
	Resnet18 + MIL	0.71	0.76	0.82	0.72
AUROC	Resnet18	0.64	0.78	**0.85**	0.7
	Resnet18 + MIL	0.66	0.65	0.76	0.62



Extended Figure 1: Slide-level aggregation using non-deep-learning methods. (a) Aggregation by majority vote, (b) Aggregation using two machine learning based classifiers, followed by a weighted sum to ensemble as the final prediction (Cao et al., 2020).



Extended Figure 2: Process of model training: training loss (left) and validation accuracy (right).



Extended Figure 3: Performance of Resnet18: (a) ROC curves with and without MIL training, (b) confusion matrix of best model, and confusion matrix on tile-level training (c) without MIL and (d) with MIL.

References

Gabriele Campanella, Matthew G. Hanna, Luke Geneslaw, Allen Miraflor, Vitor Werneck Krauss Silva, Klaus J. Busam, Edi Brogi, Victor E. Reuter, David S. Klimstra, and Thomas J. Fuchs. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nature Medicine*, 25(8):1301–1309, August 2019. ISSN 1078-8956, 1546-170X. doi: 10.1038/s41591-019-0508-1. URL http://www.nature.com/articles/s41591-019-0508-1.

Rui Cao, Fan Yang, Si-Cong Ma, Li Liu, Yu Zhao, Yan Li, De-Hua Wu, Tongxin Wang, Wei-Jia Lu, Wei-Jing Cai, Hong-Bo Zhu, Xue-Jun Guo, Yu-Wen Lu, Jun-Jie Kuang, Wen-Jing Huan, Wei-Min Tang, Kun Huang, Junzhou Huang, Jianhua Yao, and Zhong-Yi Dong. Development and interpretation of a pathomics-based model for the prediction of microsatellite instability in Colorectal Cancer. *Theranostics*, 10 (24):11080–11091, September 2020. ISSN 1838-7640. doi: 10.7150/thno.49864. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7532670/.

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, January 2021. ISSN 2072-6694. doi: 10.3390/cancers13030391. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7864494/.

Jakob Nikolas Kather, Alexander T. Pearson, Niels Halama, Dirk Jäger, Jeremias Krause, Sven H. Loosen, Alexander Marx, Peter Boor, Frank Tacke, Ulf Peter Neumann, Heike I. Grabsch, Takaki Yoshikawa, Hermann Brenner, Jenny Chang-Claude, Michael Hoffmeister, Christian Trautwein, and Tom Luedde. Deep learning can predict microsatellite instability directly from histology in gastrointestinal cancer. *Nature medicine*, 25 (7):1054–1056, July 2019. ISSN 1078-8956. doi: 10.1038/s41591-019-0462-y. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7423299/.

Eduardo Vilar and Stephen B. Gruber. Microsatellite instability in colorectal cancer—the stable evidence. *Nature Reviews Clinical Oncology*, 7(3):153–162, March 2010. ISSN 1759-4774, 1759-4782. doi: 10.1038/nrclinonc.2009.237. URL http://www.nature.com/articles/nrclinonc.2009.237.