Biomarker Prediction in Colorectal Cancer Using Multiple Instance Learning

Miljana Shulajkovska*
miljana.sulajkovska@ijs.si
Jožef Stefan Institute
Ljubljana, Slovenia

Matej Jelenc
jelenc11matej@gmail.com
Jožef Stefan Institute
Ljubljana, Slovenia

Jitenndra Jonnagaddala
jitendra.jonnagaddala@unsw.edu.au
School of Population Health, Faculty of Medicine and Health
Syndey, Australia

Anton Gradišek
anton.gradisek@ijs.si
Jožef Stefan Institute
Ljubljana, Slovenia

Abstract

Microsatellite instability (MSI) is a crucial biomarker in colorectal cancer, guiding personalised treatment strategies. The focus of our paper is on evaluating how different state-of-the-art pretrained artificial intelligence models perform in extracting features on molecular and cellular oncology (MCO) study dataset to predict biomarkers. In this study, we present an advanced approach for MSI prediction using multiple instance learning on whole slide images. Our process begins with comprehensive preprocessing of WSIs, followed by tessellation, which breaks down large images into manageable tiles. State-of-the-art feature extraction techniques are utilised on these selected tiles, employing pretrained models to capture rich, discriminative features. Various aggregation methods are applied to combine these features, leading to the prediction of MSI status across the entire slide. We assess the performance of different pretrained models within this framework, demonstrating their effectiveness in accurately predicting MSI, with results showing an AUROC of 0.91 on the MCO dataset. Our findings underscore the potential of multiple instance learning-based approaches in enhancing biomarker prediction in colorectal cancer, contributing to more targeted and effective treatment strategies.

Keywords

multiple instance learning, whole slide images, colorectal cancer, biomarker prediction

1 Introduction

MSI is a crucial biomarker in colorectal cancer (CRC) that indicates defects in the DNA mismatch repair system, leading to a high mutation rate within tumor cells. MSI status has significant clinical implications, influencing treatment decisions, particularly the use of immunotherapy, and providing prognostic information. Traditionally, MSI is determined through laboratory tests such as PCR-based assays or immunohistochemistry (IHC) on tumor tissue samples, which require invasive biopsy procedures. However, these methods can be time-consuming, costly, and dependent on the availability of sufficient tissue samples.

Deep learning methods have emerged as a promising non-invasive alternative for MSI prediction by analysing whole slide images (WSIs) of histopathological samples. These models can detect patterns linked to MSI, eliminating the need for genetic testing. WSIs provide a comprehensive view of tumor histology, offering a faster, less invasive, and more accessible means of diagnosis.

Integrating deep learning into clinical practice can improve early MSI detection, personalise treatment, and reduce invasive procedures. WSI-based methods streamline diagnostics and enhance cancer care with accessible predictive analytics.

To manage these challenges, WSIs are often divided into smaller regions or patches. A common method to address these issues is Multiple Instance Learning (MIL) [3, 8]. Due to the vast size of WSIs, computational resources can be easily overwhelmed, making MIL an essential approach. MIL is a machine learning technique that operates on sets or "bags" of instances, where the label is assigned to the entire bag rather than individual instances. This is particularly advantageous in WSI analysis, where labels such as MSI status apply to the entire slide, which is composed of numerous smaller regions or patches.

In this context, [4] demonstrates state-of-the-art (SOTA) results in predicting MSI in colorectal cancer. Their workflow utilizes the Swin-T model on small datasets to predict MSI. First, a pretrained tissue classification model is employed to filter out non-tissue patches, followed by fine-tuning a pretrained model to classify the remaining patches. Both intra-cohort and external validation are performed. When trained on the MCO dataset (N=1065), the model achieved a mean AUROC of 0.92 ± 0.05 for MSI prediction. Similarly, [11] employs a transformer-based approach for large-scale multi-cohort evaluation, involving over 13,000 patients for biomarker prediction, achieving a negative predictive value of over 0.99 for MSI prediction. When trained and tested only on a single cohort (MCO), the model achieved an AUROC of 0.85. While [4] achieved promising results on the MCO dataset using an additional tissue classifier, we obtained comparable performance without the need for tissue classification. On the other hand, [11] used a multicentric cohort, which demands additional computational resources. In comparison to their results on the MCO dataset, we achieved a 6% improvement using a smaller dataset.

In this study, we leverage MIL to process WSIs for the prediction of MSI in CRC. By testing SOTA models on the MCO dataset, we aim to assess their performance in MSI prediction using MIL. This approach not only highlights the potential of MIL in processing complex, unannotated WSIs but also contributes

Information Society 2024, 7–11 October 2024, Ljubljana, Slovenia

https://doi.org/10.70314/is.2024.scai.9705

to the broader goal of improving biomarker prediction in CRC, ultimately supporting more personalized and effective treatment strategies.

The paper is organised as follows: Section 2 outlines the methods used in the pipeline, Section 3 provides a description of the data, Section 4 presents the results, and Section 5 discusses the findings and potential directions for future work.

2 Methods

This section outlines the pipeline for MSI prediction, as illustrated in Figure 1. The process begins with the preprocessing of WSIs, including tessellation into smaller patches. Next, SOTA pretrained models are employed to extract features from these patches. These models, trained on large and diverse datasets, capture rich and discriminative features crucial for accurate MSI prediction. Finally, aggregation techniques are applied to combine the information from the patches, enabling precise MSI status prediction for the entire slide. Each subsection provides a concise explanation of these individual processes.

2.1 Preprocessing

WSIs are first tessellated into smaller, more manageable patches to facilitate further processing. This step involves dividing the large images into smaller regions using the tiatoolbox presented in [9]. Non-informative tissue patches are removed to ensure the analysis focuses solely on relevant tissue areas.

Specifically, patches that are out of bounds—where only a portion contains actual image data and the remainder consists of padding—are discarded. Patches that consist entirely of tissue are retained for subsequent analysis. This preprocessing step ensures that only informative and relevant patches are used for feature extraction and MSI prediction.

2.2 Feature Extraction Methods

Since only WSI-level annotations are available, several pretrained feature extraction models - UNI [1], ProvGigaPath [13], Phikon [2] and CTransPath [12] - are applied to patches, removing the need for detailed patch-level labeling. These SOTA models, trained on large datasets, can capture complex and discriminative features essential for accurate biomarker prediction. The extracted feature embeddings are then used as input for the aggregation and classification stages, laying the foundation for precise MSI status prediction. For technical details about these models, see Table 1.

2.3 Aggregation Methods

After feature extraction, we apply aggregation techniques to combine patch-level features into a slide-level representation. Traditional pooling methods like max-pooling and mean-pooling provide straightforward approaches.

However, these methods are limited by their lack of trainability. In recent years, attention-based pooling or ABMIL became a popular technique that adresses this issue [6]. ABMIL assigns a weight α_i to each patch's feature vector, reflecting its importance:

$$F = \sum_{i \in P} \alpha_i f_i$$

The attention scores α_i are computed as:

$$\alpha_i = \frac{\exp(w^\top \tanh(V f_i))}{\sum_{k \in P} \exp(w^\top \tanh(V f_k))}$$

where w and V are trainable parameters.

This approach allows the model to dynamically focus on the most relevant patches, leading to more accurate MSI predictions.

Another technique similiar to attention is DSMIL [7] or a dual stream aggregator, consisting of two branches, employing both an instance classifier and a bag classifier. Let $h_i \in \mathbb{R}^{L \times 1}$ be a feature embedding, and $B = \{h_0, ..., h_n\}$ a bag of embeddings. The first stream uses an instance classifier, followed by a max-pooling operation to obtain a score $c_m(B)$ and the critical embedding h_m. The second stream aggregates the embeddings into a single bag embedding which is then passed through a bag classifier:

$$c_b(B) = W_b \sum_{i}^{n-1} U(h_i, h_m) v_i$$

Where W_b is a weight vector for classification, v_i an information vector and U is a distance measurement between an arbitrary embedding and the critical embedding:

$$U(h_i, h_m) = \frac{\exp(\langle q_i, q_m \rangle)}{\sum_{k=0}^{n=1} \exp(\langle q_k, q_m \rangle)}$$

where is a query vector. Both q_i and v_i are calculated by:

$$q_i = W_q h_i, \quad v_i = W_v h_i, \quad i = 0, ..., n-1$$

where W_q and W_v are weight matrices. The final prediction is given by:

$$c(B) = \frac{1}{2}(c_m(B) + c_b(B))$$

The last approach for feature aggregation reviewed in this paper is TransMIL, as proposed in [10], a Transformer based aggregation method, which unlike the afore-mentioned methods, takes into account spatial information as well. By treating a bag of embeddings as a sequence of tokens, TransMIL uses a novel TPT module made up of two Transformer layers and a position encoding layer, where Transformer layers are designed for aggregating morphological information and Pyramid Position Encoding Generator (PPEG) which encodes spatial information, followed by a multi-layer perceptron (MLP) which classifies the bag.

2.4 MSI Classification

The aggregation step produces a single feature vector F, which encapsulates the most informative characteristics of the entire slide. This aggregated feature vector F is then passed through one or more fully connected (dense) layers. These layers apply learned weights and biases to the features to transform them into a form that is more suitable for classification. The output of the fully connected layer is often passed through an activation function, such as a sigmoid or softmax, depending on whether the classification task is binary (microsatellite instability MSI vs. microsatellite stability MSS) or multi-class. For MSI prediction, a sigmoid function is typically used, outputting a probability value between 0 and 1. The final output of the model is a single probability value indicating the likelihood of the slide being MSI. A threshold (e.g., 0.5) is applied to this probability to make a binary decision.

3 Data

For this paper the MCO study [5] was used for training and testing. The MCO study collection contains 1,500 digitized whole slide images (WSIs) of colorectal cancer tissues. Conducted by the Molecular and Cellular Oncology (MCO) Study group from 1994 to 2010, this study systematically gathered tissue samples

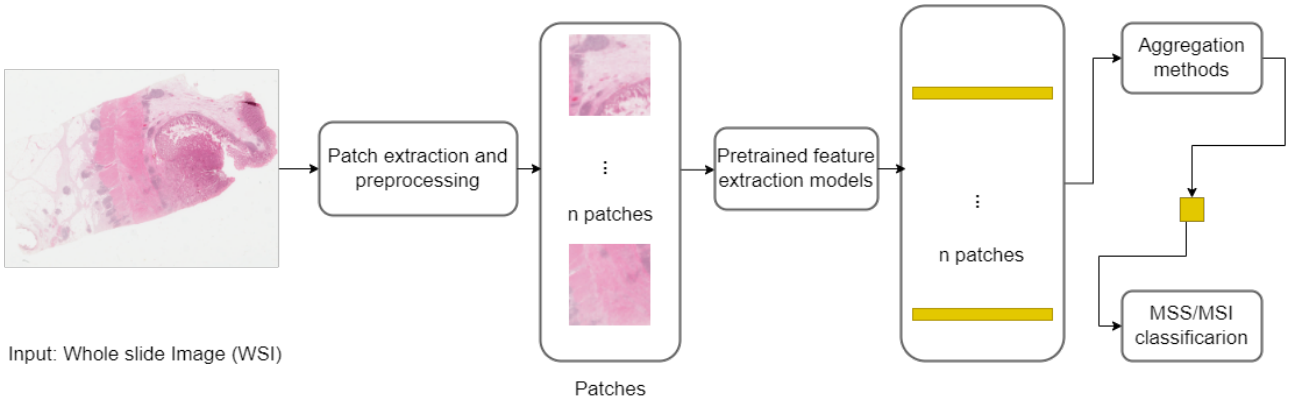


Figure 1: General architecture: multiple-instance learning approach.

feature extractor	architecture	dataset	embedding size
UNI [1]	ViT-large, DINOv2, 16 heads	Mass-100k: in-house histopathology slides from MGH and BWH, and external slides from the GTEx consortium containing >100M images, derived from >100,000 WSIs across 20 major tissue types	1024
ProvGigaPath [13]	ViT-large, DINOv2, 24 heads	Prov-Path: dataset from Providence, a large US health network comprising 28 cancer centres, consisting of 1,3B images from 171,189 WSIs	1536
Phikon [2]	ViT-large, iBOT combining MIM and CL	PanCancer40M: dataset from TCGA, covering 13 anatomic sites and 16 cancer subtypes, consisting of 43,4M images from 6,093 WSIs	768
CTransPath [12]	CNN with multi-scale Swin Transformer	dataset from TCGA and PAIP, consisting of 15M images from 32,220 WSIs	768

Table 1: Technical details about the pretrained feature extraction models.

and clinical data from over 1,500 patients who underwent colorectal cancer surgery. Each slide, representing a typical tumor section, is stained with Hematoxylin and eosin and scanned at a 40x objective, achieving a resolution of 0.25 mpp comparable to an optical microscope (~100,000 dpi). The total data size is approximately 3 Terabytes, and the collection is available on the Intersect Australia RDSI Node.

4 Results

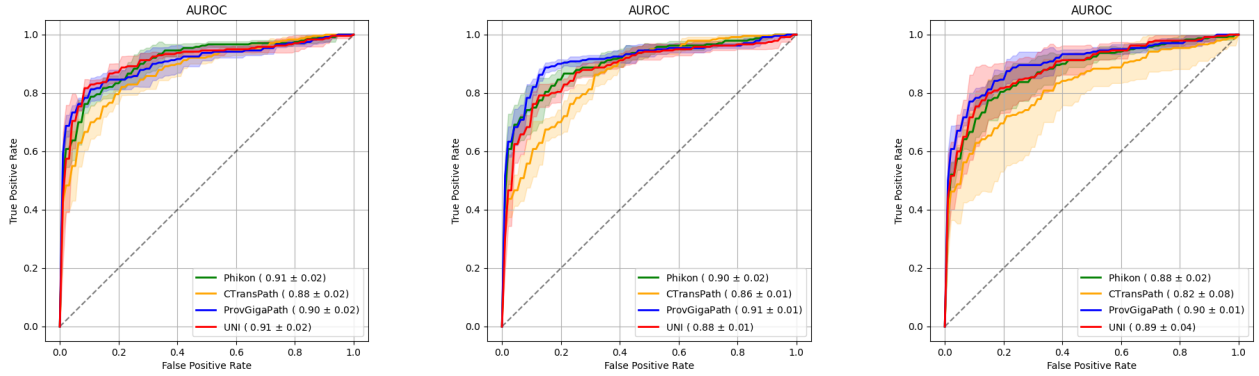
The dataset used in this study comprised 996 whole slide images (WSIs), with 242 labeled as MSI and 754 as MSS. To evaluate the performance of various aggregation methods, models were trained using 5-fold cross-validation, which ensured robust training and validation. To create a balanced testing set of 96 samples, 20% of positive (MSI) samples and an equal number of negative (MSS) samples were randomly excluded. The remaining data was split into five equally balanced parts for cross-validation, with each fold consisting of 180 samples in the validation set and 720 samples in the training set.

WSIs were then preprocessed into bags, each containing approximately 2,000 to 4,000 patches. Each patch was then converted into feature embeddings using four different feature extraction methods: Phikon, CTransPath, ProvGigaPath, and UNI. Specifically, CTransPath and Phikon produced embeddings with 768 features, UNI with 1024 features, and ProvGigaPath with 1536 features.

Three feature aggregation methods—ABMIL, DSMIL, and TransMIL—were applied to the extracted features to generate a single representative feature for each WSI. Following aggregation, a simple neural network with a sigmoid activation function and a threshold of 0.5 was used to classify MSI and MSS.

Each aggregation model was then trained for each feature extraction method on each fold, with training being conducted over 50 epochs using the AdamW optimiser and the 1-cycle learning rate scheduler to adjust the learning rate as models approached convergence. Binary cross-entropy (BCE) was used as the loss function. After each epoch, model performance was evaluated on the validation set using the AUROC metric to select the best checkpoint, as most models tended to overfit toward the end of training. The selected checkpoints were then tested to calculate the mean AUROC across all folds.

Results are presented in Figure 2a. The best performance was achieved using the DSMIL aggregation method with the ProvGigaPath feature extractor, yielding an AUROC of 0.91 ± 0.01. The ABMIL method performed best with the Phikon and UNI extractors, achieving AUROCs of 0.91 ± 0.02. Finally, the TransMIL method combined with ProvGigaPath resulted in an AUROC of 0.90 ± 0.01. Additionally, statistical analysis was performed, specifically, the Wilcoxon signed-rank test, which yielded an average p-value of 0.446, showing a relatively insignificant difference in performance of different feature extraction methods, as expected.



(a) ABMIL **(b) DSMIL** **(c) TransMIL**

Figure 2: Predictive performance of 5-fold cross-validation of different feature extractors and aggregation methods. AUROC plots for prediction of MSI/MSS status. The true positive rate represents sensitivity and the false negative rate represents 1-specificity. The shaded areas represent the standard deviation (SD). The value of the lower right each plot represents mean AUROC ± SD.

5 Discussion and Conclusion

n this study, we explored the potential of MIL combined with SOTA pretrained models for predicting MSI in colorectal cancer. Our results indicate that the approach is highly effective, achieving an AUROC of 0.913 on the MCO dataset. This is a notable achievement, particularly when compared to previous studies, such as [4] and[11], which reported AUROCs of 0.92 and 0.85, respectively, on the same dataset. Our results not only validate the effectiveness of our approach but also suggest that the careful selection and combination of feature extraction and aggregation methods can yield improvements in predictive accuracy.

The positive and negative rates observed in our results reflect the model's ability to correctly classify MSI and MSS cases. A high true positive rate (sensitivity) indicates the model's proficiency in identifying MSI-positive cases, which is crucial for ensuring that patients who could benefit from MSI-targeted therapies are accurately identified. Conversely, a high true negative rate (specificity) shows the model's effectiveness in correctly classifying MSS cases, thereby minimising false positives. To further enhance the accuracy and reliability of MSI prediction, several avenues for future work are planned.

Utilisation of the Entire Dataset: We plan to leverage the full dataset to improve the robustness of our model. Training on a larger dataset may help in capturing more nuanced patterns and variations, leading to even more accurate predictions.

Fine-Tuning of Pretrained Models: While we used pretrained models without fine-tuning in this study, fine-tuning these models specifically for the task of MSI prediction could further improve their performance. Tailoring the models to our specific data distribution and task requirements may yield significant gains in accuracy.

Incorporation of a Tissue Classifier: Since MSI is typically found in tumor tissue, we plan to integrate a tissue classifier to automatically remove non-tumor tissue from the analysis. This step should enhance the model's focus on relevant tissue regions, potentially improving MSI prediction accuracy and speed up the whole process.

Development of Advanced Aggregation Methods: We also plan to explore more sophisticated aggregation techniques that can better capture the complex relationships between patches within a WSI. Advanced methods may help in refining the prediction process, leading to further improvements in model performance.

Overall, our study demonstrates the potential of MIL-based approaches in enhancing biomarker prediction in colorectal cancer, paving the way for more personalized and effective treatment strategies.

References

[1] Richard J Chen et al. 2024. Towards a general-purpose foundation model for computational pathology. *Nature Medicine*, 30, 3, 850–862.

[2] Alexandre Filiot, Ridouane Ghermi, Antoine Olivier, Paul Jacob, Lucas Fidon, Alice Mac Kain, Charlie Saillard, and Jean-Baptiste Schiratti. 2023. Scaling self-supervised learning for histopathology with masked image modeling. *medRxiv*, 2023–07.

[3] Michael Gadermayr and Maximilian Tschuchnig. 2024. Multiple instance learning for digital pathology: a review of the state-of-the-art, limitations & future potential. *Computerized Medical Imaging and Graphics*, 102337.

[4] Bangwei Guo, Xingyu Li, Jitendra Jonnagaddala, Hong Zhang, and Xu Steven Xu. 2022. Predicting microsatellite instability and key biomarkers in colorectal cancer from h&e-stained images: achieving sota predictive performance with fewer data using swin transformer. *arXiv preprint arXiv:2208.10495*.

[5] Nick Hawkins. 2015. MCO study whole slide image collection. (2015).

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

[7] Bin Li, Yin Li, and Kevin W Eliceiri. 2021. 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*, 14318–14328.

[8] Oded Maron and Tomás Lozano-Pérez. 1997. A framework for multiple-instance learning. *Advances in neural information processing systems*, 10.

[9] Johnathan Pocock et al. 2022. Tiatoolbox as an end-to-end library for advanced tissue image analytics. *Communications medicine*, 2, 1, 120.

[10] Zhuchen Shao, Hao Bian, Yang Chen, Yifeng Wang, Jian Zhang, Xiangyang Ji, et al. 2021. Transmil: transformer based correlated multiple instance learning for whole slide image classification. *Advances in Neural Information Processing Systems*, 34, 2136–2147.

[11] Sophia J Wagner et al. 2023. Transformer-based biomarker prediction from colorectal cancer histology: a large-scale multicentric study. *Cancer Cell*, 41, 9, 1650–1661.

[12] Xiyue Wang, Sen Yang, Jun Zhang, Minghui Wang, Jing Zhang, Wei Yang, Junzhou Huang, and Xiao Han. 2022. Transformer-based unsupervised contrastive learning for histopathological image classification. *Medical image analysis*, 81, 102559.

[13] Hanwen Xu et al. 2024. A whole-slide foundation model for digital pathology from real-world data. *Nature*, 1–8.