Computer Methods and Programs in Biomedicine 280 (2026) 109317

Contents lists available at ScienceDirect

Computer Methods and Programs in Biomedicine

journal homepage: www.sciencedirect.com/journal/computer-methods-and-programs-in-biomedicine

dMMR prediction from colorectal cancer histopathology: Leveraging non-tumor and low-magnification regions

Liisa Petäinen [a,*], Juha P. Väyrynen [b], Jan Böhm [c], Pekka Ruusuvuori [d,e], Maarit Ahtiainen [f], Hanna Elomaa [g], Henna Karjalainen [b], Meeri Kastinen [b], Vilja V. Tapiainen [b], Ville K. Äijälä [b], Päivi Sirniö [b], Anne Tuomisto [b], Markus J. Mäkinen [b], Jukka-Pekka Mecklin [g,h], Ilkka Pölönen [a], Sami Äyrämö [a,i]

[a] Faculty of Information Technology, University of Jyväskylä, Finland. Address: Mattilanniemi 2, 40100 Jyväskylä, Finland
[b] Translational Medicine Research Unit, Medical Research Center, Oulu University Hospital and University of Oulu, Oulu, Finland
[c] Department of Pathology, Wellbeing Services County of Central Finland, Jyväskylä, Finland
[d] Cancer Research Unit, Institute of Biomedicine, University of Turku, Turku, Finland
[e] Faculty of Medicine and Health Technology University of Tampere, Tampere, Finland
[f] Central Finland Biobank, Wellbeing Services County of Central Finland, Jyväskylä, Finland
[g] Department of Education and Research, Wellbeing Services County of Central Finland, Jyväskylä, Finland
[h] Faculty of Sport and Health Sciences, University of Jyväskylä, Jyväskylä, Finland
[i] Wellbeing Services County of Central Finland, Jyväskylä, Finland

ARTICLE INFO

Keywords:
Colorectal cancer
DNA mismatch-repair deficiency
Microsatellite instability
MMR
MSI
Multi-scale
Deep learning
Digital pathology
Artificial intelligence
Histopathology, foundation model

ABSTRACT

Background and Objective: Colorectal cancer is the second leading cause of cancer-related mortality worldwide, posing a substantial burden on healthcare systems. Identifying DNA mismatch repair deficiency (dMMR) is critical for guiding treatment, yet conventional methods rely on labor-intensive DNA analysis. While deep-learning approaches have shown promise for predicting dMMR from histopathological images, most studies focus exclusively on tumor regions and single-scale representations. This study systematically evaluates the predictive value of tumor and non-tumor regions across multiple magnifications for dMMR prediction from whole-slide images (WSIs).

Methods: A total of 24 different modeling approaches were evaluated, varying by tissue origin (tumor vs. non-tumor), magnification level (5x and 20x), and tile embedding strategy, including digital pathology foundation models. Tile embeddings were further trained with 1228 WSIs using multiple-instance learning (MIL) based approach. The best-performing configurations were selected for external evaluation. External testing was carried out on two independent cohorts consisting of 1010 and 457 WSIs, respectively.

Results: Non-tumorous regions demonstrated measurable predictive value, although performance remained lower than that obtained from tumor regions (F1 = 0.896, precision = 0.888, sensitivity = 0.594, specificity = 0.982). Among the nine models selected during internal validation, the top three models—one multi-scale approach and two models trained on 20x tumor regions—achieved F1 scores of 0.870–0.889 with precision of 0.885–0.920, sensitivity of 0.852, and specificity of 0.889–0.926. On external validation, the top three models, all based on foundation-model tile embeddings, achieved F1 scores of 0.916–0.919 on the first cohort and 0.928–0.934 on the second cohort. Across cohorts, specificity remained consistently high (0.964–0.992), while sensitivity ranged from 0.500 to 0.682.

Conclusion: This study demonstrates that dMMR status in colorectal cancer can be effectively predicted from histopathological WSIs using MIL-based models, with moderate generalizability across independent cohorts. In addition to confirming the predictive value of tumor regions, the results reveal that non-tumorous tissue also contains detectable predictive signals, suggesting that microenvironmental features may contribute to dMMR-associated histological patterns. Furthermore, the use of foundation model–derived embeddings improved generalizability across datasets. Future work should explore integrating non-tumor tissue features and clinical data to further improve predictive performance.

* Corresponding author.
E-mail address: liisa.h.petainen@jyu.fi (L. Petäinen).

https://doi.org/10.1016/j.cmpb.2026.109317
Received 17 September 2025; Received in revised form 11 March 2026; Accepted 13 March 2026
Available online 17 March 2026

1. Introduction

Artificial intelligence (AI) has become increasingly influential in medicine, with machine learning systems demonstrating strong performance in tasks such as diagnosis, risk prediction, workflow optimization, and precision medicine [1–3]. Despite the active research and technical development of AI in digital pathology and medicine in general, adopting AI into routine clinical workflow remains limited [4,5]. Within digital pathology, deep learning models now extract quantitative morphological patterns from routine H&E slides and can predict clinically relevant features such as tumor subtypes, molecular alterations, and patient outcomes [1,5].

One of these molecular alterations is the deficiency of the DNA mismatch repair (dMMR) mechanism, which often leads to microsatellite instability (MSI), and is observed in 10–15 % of all colorectal cancer (CRC) cases [6]. Detecting dMMR is clinically important as dMMR tumors are predicted to respond well to immunotherapy, and might be a sign of a hereditary type of CRC, the Lynch syndrome. Even though MMR testing is recommended for all CRC patients [7], it is not always performed due to the associated labour and material costs. The most frequently applied methods for defining the MMR status are immunohistochemistry (IHC), polymerase chain reaction (PCR), and next-generation sequencing (NGS). In the past few years, deep learning has shown its potential to predict dMMR straight from standard hematoxylin-eosin (H&E) stained specimens in CRC and other cancer types [8]. It is estimated that AI-assisted, cloud-based dMMR screening would offer a cost-saving advantage of approximately 370-fold compared to conventional IHC panels and 580- fold savings compared to NGS, highlighting AI's potential as a highly economical alternative for dMMR detection [9]. AI could reduce turnaround time providing results in less than one day in contrast to the approximately 4-day processing required for IHC and the 12-day timeframe for NGS, thereby facilitating faster clinical decision-making and enabling earlier treatment initiation. Combining AI with a confirmatory approach, such as PCR or IHC, was estimated to bring the best cost savings without significant loss of diagnostic accuracy [9].

A common approach of AI-assisted dMMR prediction in CRC is to train the model with image patches tiled from the tumor regions, often with a magnification of 20x [10–13]. Additionally, incorporating mucinous patches in the training data has also demonstrated satisfactory performance [14]. While dMMR models typically achieve high sensitivity (>90 %), their specificity needs improvement. For example, in experiments by Echle et al. [11], the specificity varied from 19 % to 79 %. By adjusting factors such as the operating threshold, the choice of model can be tailored to clinical needs, with a focus on either high sensitivity or high specificity. Comparing the results from previous studies is challenging as specificity and other statistical details are not always reported, complicating the evaluation of model performance across different approaches and datasets.

Deep learning models typically learn patterns from the training data and a common challenge in the field is the lack of generalizability. A dMMR detection model exhibiting a high level of generalizability was developed using WSIs from multiple countries, gaining AUROC of 0.95 on both internal and external test sets [11]. The most commonly used dataset for training dMMR models in CRC is colon and rectal cancer WSIs from The Cancer Genome Atlas [8]. In addition to CRC, dMMR detection models have been developed for various cancer types, including endometrial cancer, where the clinical significance of dMMR is comparable to that in CRC [8]. However, the performance of these models in other cancer types has not matched that observed in CRC, likely due to the comparatively limited availability of data.

Being a high-resolution image, WSIs are tiled before training. The most commonly applied method in dMMR models involves training a convolutional neural network (CNN) model with tumor patches, particularly using ResNet-based architectures, such as ResNet18 [15]. This architecture is favoured for its relatively small number of parameters and effectiveness with histopathological images [8,12]. To obtain patient-level predictions from patch-level probabilities, several strategies have been implemented. These methods include majority voting, computation of the mean or median of patch-wise probabilities (with predetermined thresholds), using Multiple-Instance learning (MIL) strategies, and more complex pipelines [10,13].

Combining patches extracted from different magnification levels to develop a multi- scale model has been applied to tasks like classification [16–18], segmentation [19–22], recurrence prediction [23] and detection [24,25]. A multi-scale approach has been used for predicting molecular changes, as seen in a study by Jain et al. [26] predicting mutational burden from lung cancer WSIs, where the multi-scale (5x, 10x, 20x) model outperformed single-scale models. Cao et al. [27], however, compared three magnification levels for dMMR detection and found the best performance with a 20x magnification.

Although multi-scale approaches have been explored in computational pathology, most prior studies have focused primarily on tumor regions and have not systematically evaluated the predictive contribution of non-tumorous tissue or compared different magnification levels of tumor regions within the same experimental framework. Histopathological predictors of dMMR include non-tumorous features such as tumor-infiltrating lymphocytes [28] and Crohn's-like inflammatory reactions [29], suggesting that informative signals may also be present outside tumor regions.

In this study, we investigate the predictive value of both tumor and non-tumor regions across different magnification levels for dMMR prediction in CRC. Furthermore, we evaluate multiple tile embedding strategies, including recently proposed digital pathology foundation models, and assess model generalizability across large independent external cohorts. In addition to providing architectural context, the use of lower magnification levels may reduce computational complexity, which could facilitate the future integration of digital pathology workflows into clinical practice.

2. Materials and methods

2.1. Data

A dataset consisting of 1282 WSIs (CRC-FFPE-FIN) was applied to the model development in this study. Of these WSIs, 152 WSIs were classified as sporadic dMMR, 112 as hereditary dMMR, and 1018 as pMMR (MMR proficient). The samples are collected from multiple hospital centres in Finland. CRC-FFPE-FIN-data consists of WSIs from patients with primary colorectal cancers (stages I-IV), with one representative tumor WSI from each patient. All WSIs were scanned with Hamamatsu NanoZoomer-XR with a resolution of 0.5 μm per pixel (MPP). The MMR status was defined as described in Elomaa et al. [30].

Before starting the training and validation phases, 27 dMMR and 27 pMMRs from CRC-FFPE-FIN samples were put aside from all the processing to be used later as internal test data. Thereafter, the remaining ($N = 1228$) WSIs were split in 1:2 ratio into training and validation sets. A diagram of the data flow is presented in Fig. 1. For external testing, two datasets were used: one from Oulu University Hospital, comprising 1010 WSIs (CRC—OYS-FIN), and a second dataset with colon and rectum WSIs from the public TCGA database. The MMR status in CRC—OYS-FIN was determined using the same approach as for CRC-FFPE-FIN data. In the TCGA dataset, MMR status was determined based on the study by Liu et al. [31], where MSI-high samples represented the dMMR group, while MSI-low and microsatellite stable (MSS) samples were classified as pMMR. Summary of WSI scanning parameters for each dataset is shown in Table 1.

2.2. Preprocessing

Each WSI in the training group was tiled into three training subgroups: tumor patches with a magnification of 20x (TUM20x), tumor

patches with a magnification of 5x (TUM5x) and patches from non-tumorous sites, i.e., subgroup ”other”, with a magni- fication of 5x (OTHER5x). Example patches from each subgroup are shown in Fig. 2.

The tiling was accomplished as follows. At first, the tumor parts and parts belonging to group “other” were detected, masked and tiled using a deep learning model developed for automated scoring of tumor-stroma ratio [32]. The image patches were tiled using a sliding window procedure and the current patch was tiled if 75 % of the 5x window was within the tumor/other mask. Tiling was done at magnification levels 5x and 20x, and the patches in both magnification levels shared the same centre point. In group other, the tiling was done at a magnification level of 5x.

As WSIs exhibit variations in the size of tumor tissue, the quantity of patches extracted from each WSI varies significantly. Moreover, given that dMMR is a less prevalent characteristic compared to pMMR, the number of WSIs demonstrating pMMR was approximately five times greater than those exhibiting dMMR within the training set. To harmonize the dataset while preserving the heterogeneity inherent in the larger subset of pMMR WSIs, up to 300 patches were randomly sampled from each pMMR WSI and up to 1500 from each dMMR WSI. This approach enabled the utilization of all pMMR WSIs while maintaining data volumes within specified limits and ensuring the training data remained as balanced as possible. The number of WSIs and patches in the training and validation set is shown in Fig. 1. All the patches were color normalized using Macenko’s method [33] and normalized tiles were visually inspected to ensure that no artefacts or distortions were introduced during preprocessing.

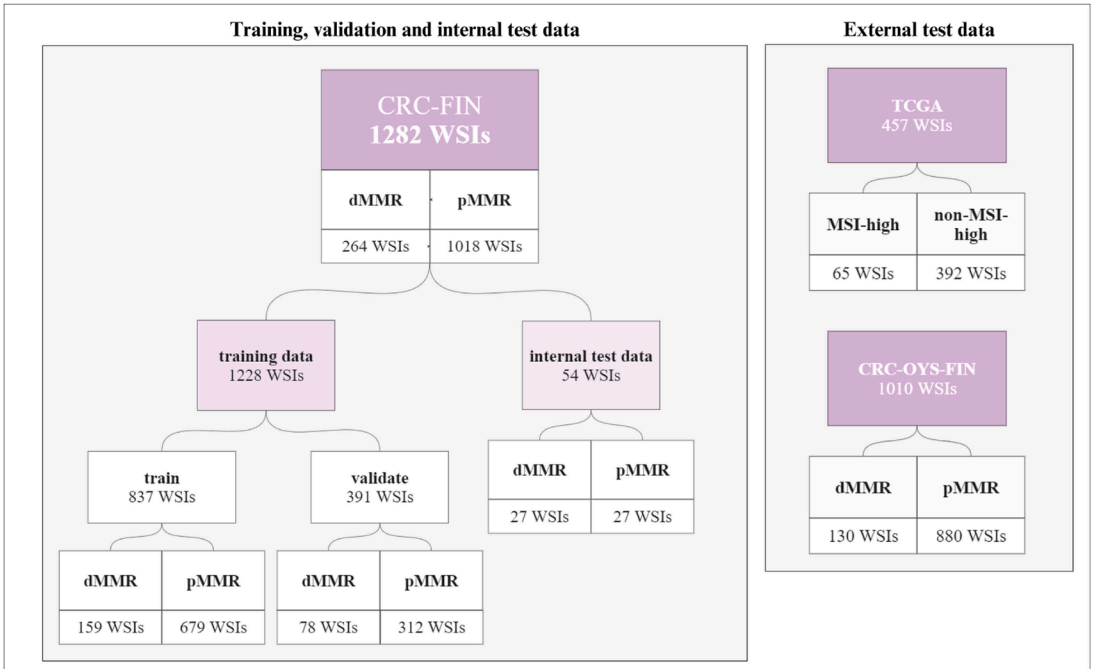
Table 1
Summary of WSI scanning parameters for each dataset. The scanning magnifications of the CRC—OYS-FIN dataset were either 20x or 40x, with the majority of slides scanned at 20x magnification. The magnification was verified for each file individually to ensure that the tiling area remained consistent across all images.

Dataset	MPP	Scanning magnification	Scanner
CRC-FFPE-FIN	0.5	20x	Hamamatsu NanoZoomer-XR
CRC—OYS-FIN	0.5 or 0.25*	20x or 40x*	Hamamatsu NanoZoomer-S360
TCGA	0.5	20x	Aperio scanners

2.3. Training and validation of the models

The model is developed using aggregated tile-level embeddings, with multiple methodological variants at each stage of the pipeline. The workflow comprises three main steps: 1) Tile-level feature extraction from tumor regions at two magnifications (TUM5x and TUM20x) and from non-tumor regions (OTHER5x), 2) feature aggregation and slide-level classification.

Tile-level feature extraction (step 1) is conducted using three distinct approaches: a lightweight MobileNetV3 [34] trained on the



The number of WSIs and patches in training and validation data.

	Train		Validation	
	dMMR	pMMR	dMMR	pMMR
WSIs	159	679	78	312
Patches TUM5x[1]	158,712	180,616	65,665	81,714
Patches OTHER5x	234,722	203,700	113,159	177,502

[1]Since TUM5x and TUM20x share the same centre point, the quantity of TUM20x patches is identical to that of TUM5x patches.

Fig. 1. Overview of the data flow in this study. At first, a fraction of the entire dataset was extracted as internal test data. The training data was split into train and validation. In addition to the internal test set, two external test data sets were utilized to assess the generalization performance of the individual and multi-scale models chosen in the validation phase. The table at the bottom of the figure describes the number of WSIs and patches in training and validation data.

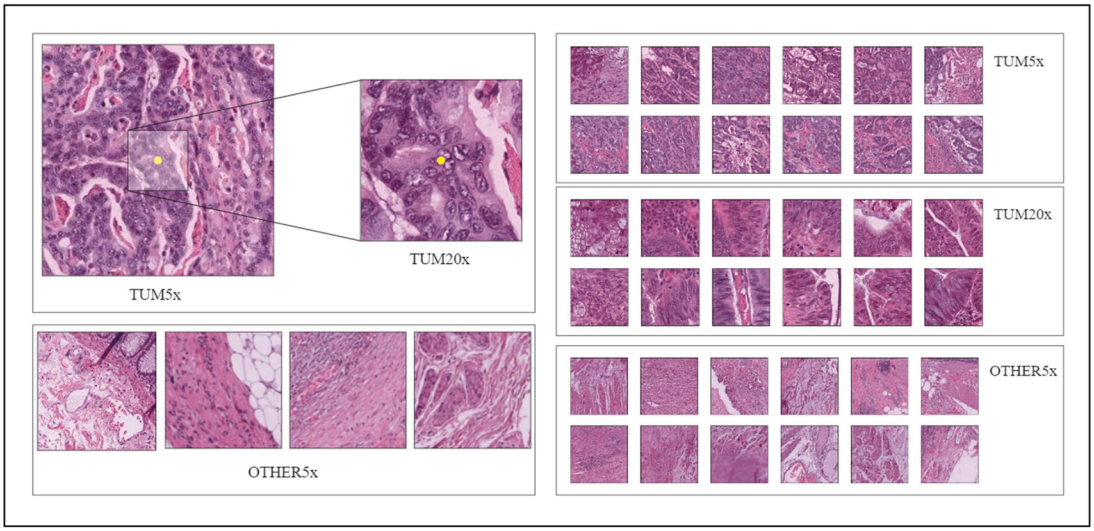


Fig. 2. Tumor regions were detected and tiled from 5x and 20x magnification levels using the same centre point. All tiles were resized to 224 × 224 pixels. Non-tumorous regions were tiled using 5x magnification. Example tiles from each subgroup are shown on the right side of the figure.

CRC-FFPE-FIN training cohort (837 WSIs), a digital pathology foundation model UNI-v2 [35], and another foundation model HOptimus0 [36]. MobileNetV3 was based on tile-level accuracy comparison and training time between four architectures (Supplemental Table 1). The MobileNetV3 models were initialized with ImageNet-pretrained weights [37] and trained separately for each subgroup (TUM5x, TUM20x, and OTHER5x) of tiles. Each tile inherited the slide-level pMMR/dMMR label as ground truth, as defined in Section 2. The hyperparameters were chosen using 5-fold cross-validation (Supplemental Table 2). The models were then trained using the best-performing hyperparameters and the optimal number of epochs was determined using an early-stopping model selection strategy.

For feature aggregation and final classification (step 2), two methods are examined: Attention-based Deep Multiple Instance Learning (ABMIL) originally presented by Ilse et al., [38] and clustering-constrained-attention multiple-instance learning (CLAM), specifically designed for digital pathology by Lu et al. [39]. The models with single scale are trained with CRC-FFPE-FIN data using slide-level labels as a ground truth.

Multi-scale training employed two parallel MIL branches for each magnification, optimized end-to-end and fused by concatenating their slide-level embeddings before classification layer. Training was regularized using an early stopping strategy based on validation performance.

For experiments utilizing MobileNetV3-derived tile embedding, the final model in step 2 is trained on the validation subset of CRC-FFPE-FIN (391 WSIs), partitioned into train/validation splits using a 70/30 ratio. In experiments where tile embeddings are derived from UNI-v2 or HOptimus0, the combined train+validation cohort of CRC-FFPE-FIN (1228 WSIs) is used for training the final model, employing the same 70/30 split. The overall training configuration and multi-scale model structure are illustrated in Fig. 3.

The optimal combination of tissue, magnification and tile embedding origin and MIL-approach is selected based on comparative evaluation across all configurations and chosen models are subsequently assessed using both internal and external test sets.

2.4. Testing the validated models

The WSIs in both internal and external testing groups were tiled following the same methods as with the training and validation WSIs, all tiles from each WSI and each subgroup (TUM5x, TUM20x) were included. If fewer than 30 tumor tiles were obtained from a WSI, it was extracted from the test set. Due to the imbalanced class distribution the

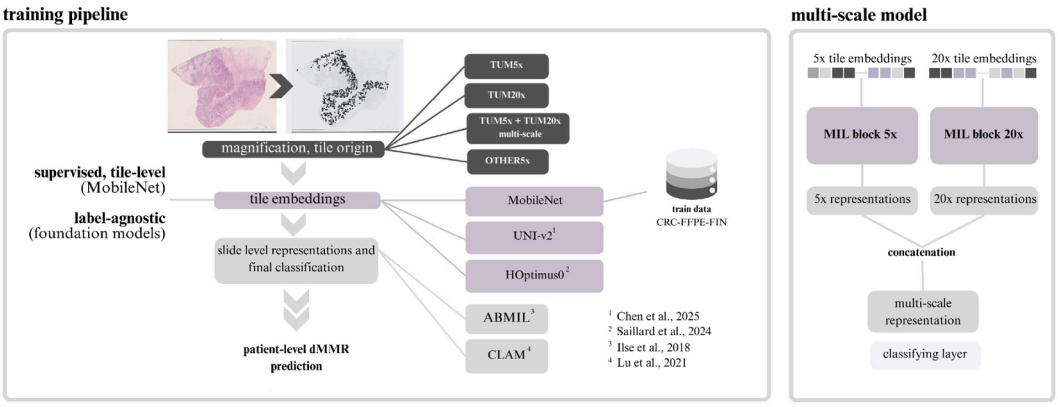


Fig. 3. Overview of the training pipeline (left) and multi-scale model structure (right). Tile-level embeddings are extracted using supervised MobileNet or label-agnostic foundation models (step 1). Tile embeddings are used as input to train the final classifier, a MIL-based model with ABMIL or CLAM architecture (step 2). The best-performing training strategy combination is then selected based on validation results. In the multi-scale model, the representations from two tumor magnifications are concatenated before the final classification layer.

key evaluation metrics were AUCPR and F_1 score. The aim was to guarantee robust and reliable detection ability for the minority class (dMMR) since it is of the greatest interest in the present task.

2.5. Equipment and software

All models were developed with Python version 3.11.13. on Linux GPU server Tesla P100, x 86 64 using PyTorch- and TorchVision-libraries, versions 2.2.2+cu118 and 0.16.0+cu121, respectively. The masking of the WSIs was performed with OpenCV 4.5.2. The Open-source library StainTools was utilized for color normalization, it is available for download at GitHub: https://github.com/Peter554/StainTools. Performance metrics were computed using the metrics module of Scikit-learn version 1.8.0.

3. Results

3.1. Validation results

Patient-level dMMR prediction classification performance on the validation set is shown in Table 2 and Area Under the Precision-Recall (AUPRC) curves in Fig. 4 for the three best performing models from all three tile feature origins (total of nine approaches, shown bold on Table 2). Boxplots for visualizing the validation results grouped by tissue origin and magnification, and by tile feature origin are shown in Fig. 5. The validation set consisted of 78 WSIs with dMMR and 312 WSIs with pMMR (see Fig. 1).

3.2. Test results

From all 24 approaches validated, top-3 best performing models from each tile embedding origin (MobileNet, HOptimus0, UNI-v2) were chosen for testing on internal and external test sets. The results are shown in Table 3. Boxplots showing the AUPRC grouped by tissue origin and magnification, as well as by tile embedding origin are visualized in Fig. 6.

4. Discussion

This study explored the potential of a multi-scale approach and non-tumorous regions to predict dMMR in CRC. The tumor regions from 20x magnification as well as multi-scale approach showed the best performance on validation phase. Also, the low-magnification of non-tumorous regions showed rather good performance. When comparing the results of 5x and 20x single-scale tumor models, the 20x model performed slightly better in validation and internal test sets, which is consistent with the previous study by [27]. The non-tumorous regions were not as predictive as tumorous regions when analysed collectively. However, further research is needed to better understand their potential and to assess whether specific tissue types within non-tumorous regions could provide predictive information when analyzed separately. For example, among the histomorphological features analyzed, extracellular mucus was recognized as the strongest individual predictor of MSI [14]. One challenge in the multi-scale model is the inconsistent availability of sufficient tumorous areas in specimens, especially noticeable at a 5x magnification level. Exploring the potential contribution of non-tumorous areas could address this challenge.

Another challenge in the field of digital pathology and machine learning is the lack of generalizability. The tests on two external datasets showed that fine-tuning the model with laboratory-specific data is necessary for gaining satisfying performance. This is in line with other MSI/dMMR studies, which show that performance varies between internal and external test sets. According to Li et al. [40], the pooled MSI/dMMR identification performance of deep learning models on external validation sets was 0.80 for sensitivity and 0.54 for specificity. In our study, the mean sensitivity on the external test sets was 0.586 and mean specificity 0.981.

These common generalization challenges can arise for several reasons. For example, because of their high learning capacity, deep learning models may capture dataset- or site-specific patterns during training, which can limit their ability to generalize across laboratories. Variations in pre-diagnostic procedures—such as tissue fixation protocols—and differences in scanning equipment can introduce inconsistencies that make cross-laboratory application challenging. Variations in WSI scanners, slight differences in histopathology procedures across countries, and ethnic heterogeneity may explain this finding. To improve the

Table 2
Validation results of all 24 approaches. Top-3 of all three different tile feature origins (MobileNet, HOptimus0, and UNI-v2) approaches were selected for internal and external testing (shown bold). The multi-scale approach includes embeddings from two magnification levels of tumor tissue, model structure illustrated in Fig. 3. PPV = Positive predictive value, NPV = Negative predictive value.

Tile embeddings	Aggregator	Tissue origin and magnification	Weighted F1	Sensitivity	Specificity	PPV	NPV	AUPRC
MobileNet	ABMIL	TUM5x	0,955	0,864	0,978	0,905	0,967	0,992
MobileNet	CLAM	TUM5x	0,946	0,864	0,966	0,864	0,966	0,991
MobileNet	**ABMIL**	**TUM20x**	**0,991**	**0,955**	**1,000**	**1,000**	**0,989**	**0,997**
MobileNet	**CLAM**	**TUM20x**	**0,991**	**1,000**	**0,989**	**0,957**	**1,000**	**0,999**
MobileNet	multi-scale ABMIL	multi (TUM5x+TUM20x)	0,973	0,909	0,989	0,952	0,978	0,996
MobileNet	**multi-scale CLAM**	**multi (TUM5x+TUM20x)**	**0,982**	**0,955**	**0,989**	**0,955**	**0,989**	**0,999**
MobileNet	ABMIL	OTHER5x	0,939	0,783	0,979	0,900	0,948	0,983
MobileNet	CLAM	OTHER5x	0,957	0,870	0,979	0,909	0,968	0,985
HOptimus0	ABMIL	TUM5x	0,917	0,556	0,980	0,806	0,935	0,919
HOptimus0	CLAM	TUM5x	0,908	0,533	0,973	0,750	0,931	0,911
HOptimus0	**ABMIL**	**TUM20x**	**0,930**	**0,778**	**0,952**	**0,714**	**0,965**	**0,945**
HOptimus0	**CLAM**	**TUM20x**	**0,938**	**0,756**	**0,966**	**0,773**	**0,963**	**0,940**
HOptimus0	**multi-scale ABMIL**	**multi (TUM5x+TUM20x)**	**0,938**	**0,756**	**0,966**	**0,773**	**0,963**	**0,943**
HOptimus0	multi-scale CLAM	multi (TUM5x+TUM20x)	0,930	0,778	0,952	0,714	0,965	0,941
HOptimus0	ABMIL	OTHER5x	0,863	0,435	0,989	0,909	0,877	0,939
HOptimus0	CLAM	OTHER5x	0,886	0,522	0,989	0,923	0,894	0,943
UNI-v2	ABMIL	TUM5x	0,916	0,644	0,959	0,707	0,946	0,931
UNI-v2	CLAM	TUM5x	0,922	0,689	0,959	0,721	0,953	0,931
UNI-v2	**ABMIL**	**TUM20x**	**0,934**	**0,733**	**0,966**	**0,767**	**0,959**	**0,938**
UNI-v2	**CLAM**	**TUM20x**	**0,937**	**0,733**	**0,969**	**0,786**	**0,959**	**0,940**
UNI-v2	**multi-scale ABMIL**	**multi (TUM5x+TUM20x)**	**0,935**	**0,756**	**0,962**	**0,756**	**0,962**	**0,942**
UNI-v2	multi-scale CLAM	multi (TUM5x+TUM20x)	0,931	0,711	0,966	0,762	0,956	0,938
UNI-v2	ABMIL	OTHER5x	0,843	0,391	0,979	0,818	0,868	0,917
UNI-v2	CLAM	OTHER5x	0,889	0,565	0,979	0,867	0,902	0,938

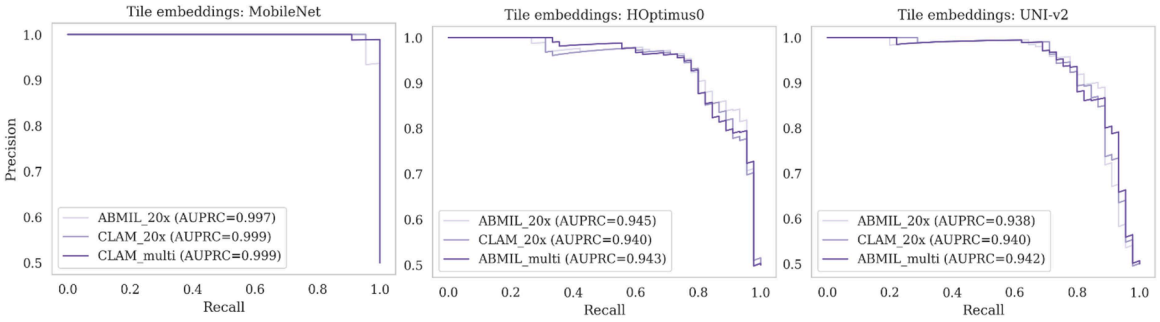


Fig. 4. AUPRC curves of three best performing models on validation set, grouped by the origin of tile embeddings.

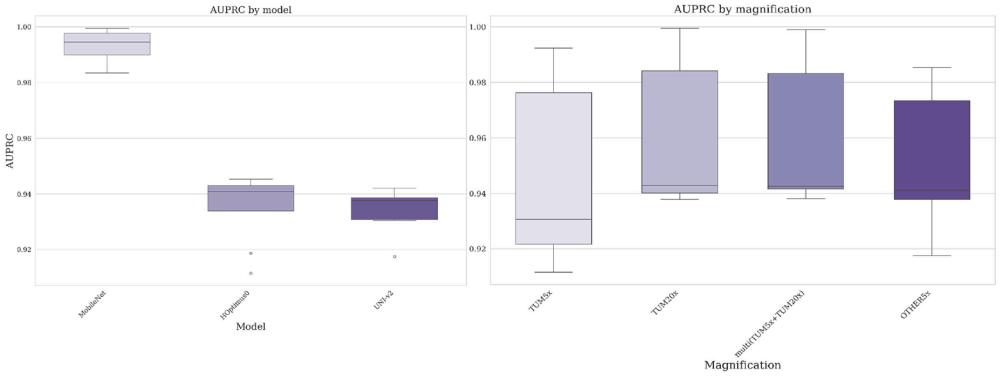


Fig. 5. Boxplots showing the AUPRC of validation results grouped by model applied for the tile embeddings (left) and by tissue origin and magnification (right).

Table 3
Test results on the nine different approaches. Top-3 best performing approaches on each test set shown bold.

Tile embeddings	Aggregator	Tissue origin and magnification	Weighted F1	Sensitivity	Specificity	PPV	NPV	AUPRC
Internal test set								
CRC-FFPE-FIN, $N = 54$ (27+27)								
MobileNet	**ABMIL**	**TUM20x**	**0,889**	**0,852**	**0,926**	**0,920**	**0,862**	**0,958**
MobileNet	**CLAM**	**TUM20x**	**0,870**	**0,852**	**0,889**	**0,885**	**0,857**	**0,953**
MobileNet	**multi-scale CLAM**	**multi (TUM5x+TUM20x)**	**0,870**	**0,852**	**0,889**	**0,885**	**0,857**	**0,935**
HOptimus0	ABMIL	TUM20x	0,832	0,741	0,926	0,909	0,781	0,950
HOptimus0	CLAM	TUM20x	0,790	0,630	0,963	0,944	0,722	0,955
HOptimus0	multi-scale ABMIL	multi (TUM5x+TUM20x)	0,790	0,630	0,963	0,944	0,722	0,957
UNI-v2	ABMIL	TUM20x	0,766	0,556	1,000	1,000	0,692	0,947
UNI-v2	CLAM	TUM20x	0,766	0,556	1,000	1,000	0,692	0,942
UNI-v2	multi-scale ABMIL	multi (TUM5x+TUM20x)	0,808	0,630	1,000	1,000	0,730	0,951
External test sets								
CRC—OYS-FIN, $N = 915$ (121+794)								
MobileNet	ABMIL	TUM20x	0,579	0,910	0,453	0,204	0,970	0,669
MobileNet	CLAM	TUM20x	0,384	0,943	0,252	0,162	0,966	0,615
MobileNet	multi-scale CLAM	multi (TUM5x+TUM20x)	0,526	0,959	0,388	0,194	0,984	0,677
HOptimus0	**ABMIL**	**TUM20x**	**0,918**	**0,533**	**0,986**	**0,855**	**0,932**	**0,957**
HOptimus0	CLAM	TUM20x	0,905	0,443	0,990	0,871	0,920	0,954
HOptimus0	multi-scale ABMIL	multi (TUM5x+TUM20x)	0,907	0,459	0,989	0,862	0,922	0,951
UNI-v2	ABMIL	TUM20x	0,912	0,475	0,991	0,892	0,925	0,950
UNI-v2	**CLAM**	**TUM20x**	**0,919**	**0,508**	**0,992**	**0,912**	**0,929**	**0,954**
UNI-v2	**multi-scale ABMIL**	**multi (TUM5x+TUM20x)**	**0,916**	**0,500**	**0,991**	**0,897**	**0,928**	**0,956**
TCGA, $N = 298$ (41+257)								
MobileNet	ABMIL	TUM20x	0,439	0,955	0,302	0,163	0,979	0,605
MobileNet	CLAM	TUM20x	0,318	1,000	0,195	0,151	1,000	0,579
MobileNet	multi-scale CLAM	multi (TUM5x+TUM20x)	0,359	0,955	0,231	0,151	0,973	0,606
HOptimus0	**ABMIL**	**TUM20x**	**0,934**	**0,636**	**0,981**	**0,824**	**0,950**	**0,945**
HOptimus0	CLAM	TUM20x	0,926	0,568	0,984	0,833	0,941	0,943
HOptimus0	multi-scale ABMIL	multi (TUM5x+TUM20x)	0,916	0,523	0,981	0,793	0,935	0,929
UNI-v2	ABMIL	TUM20x	0,927	0,659	0,968	0,744	0,952	0,931
UNI-v2	**CLAM**	**TUM20x**	**0,928**	**0,682**	**0,964**	**0,732**	**0,955**	**0,938**
UNI-v2	**multi-scale ABMIL**	**multi (TUM5x+TUM20x)**	**0,932**	**0,659**	**0,974**	**0,784**	**0,952**	**0,943**

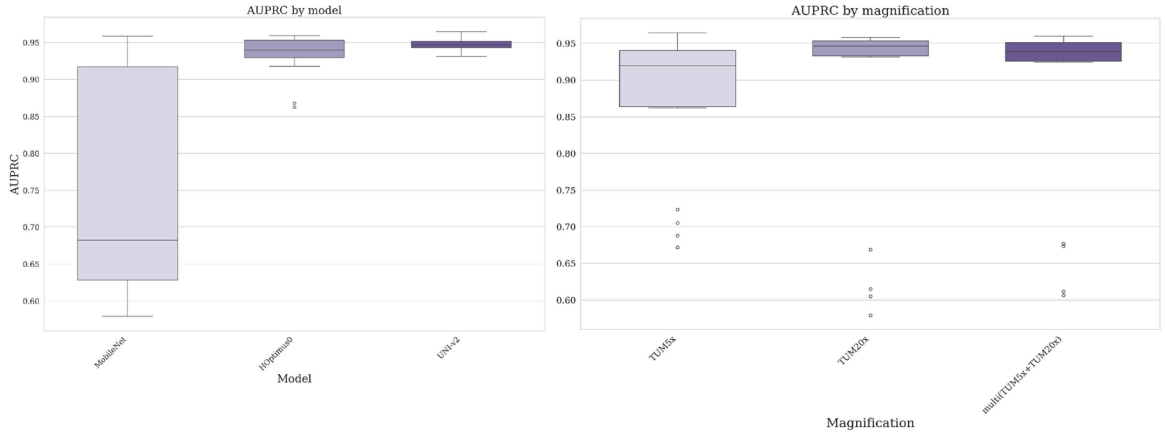


Fig. 6. Boxplots showing the AUPRC of all test results grouped by model applied for the tile embeddings (left) and by tissue origin and magnification (right).

performance, retraining or fine-tuning the MIL-layer for each dataset would likely yield better results. Also, additional clinical information could increase the predictive performance of MSI/dMMR detection models [41].

Using low-level magnification could simplify the automated dMMR detection, which is advantageous especially if the model is intended for use as a prescreening tool. Thus, model selection should effectively balance between performance and practicality to enhance utility across diverse applications. Nevertheless, the generalizability issues prevalent in digital pathology are also evident in this study, highlighting the need for domain adaptation or fine-tuning for optimal performance across different datasets.

Our study demonstrates that both magnification levels from tumorous regions, as well as low-magnification features from non-tumorous regions, independently provide predictive value for dMMR. Among all evaluated models, the multi-scale approach and as well as the model with 20x tumorous regions yielded the highest performance across both internal and external test sets.

With respect to clinical implementation, a hybrid workflow—where the computational model acts as an initial pre-screening step prior to confirmatory IHC or PCR—could meaningfully reduce laboratory burden, shorten turnaround times, and help prioritize cases requiring further molecular testing. For clinical deployment, these models should also be calibrated or fine-tuned to the institution's own data environment to ensure optimal performance and reliable results.

Glossary

CRC	colorectal cancer
CNN	convolutional neural network
dMMR	deficiency of the DNA mismatch repair
MPP	microns per pixel
MSS	microsatellite stable
MSI	microsatellite instability
pMMR	DNA mismatch repair proficient
WSI	whole slide image

Funding

This study is one part of the Central Finland AI hub II project that has received funding from the Regional Council of Central Finland (https://www.keskisuomi.fi/) and the European Regional Development Fund (ERDF) (https://ec.europa.eu/regional policy/funding/erdf en). The data (CRC samples, WSIs and MMR analysis) collected in this work was supported by Jane and Aatos Erkko Foundation (https://jaes.fi/en/frontpage/). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Ethics approval and consent to participate

CRC-FFPE-FIN: Regional medical research ethics committee of the Wellbeing services county of Central Finland (Dnro 13U/2011, 1/2016, 8/2020, 2/2023), FIMEA (FIMEA/2023/001573) and Central Finland Biobank (BB23–0172). The need to obtain informed consent from the study patients was waived (FIMEA/2023/001573). Oulu: Regional medical research ethics committee of the Wellbeing services county of North Ostrobothnia (25/2002, 42/2005, 122/2009, 37/2020), Biobank Borealis (BB- 2017 1012) and Fimea (FIMEA/2022/001941). The participants gave written informed consent for the study.

Data availability

All WSIs from the TCGA-COAD and TCGA-READ are available from the Genomic Data Commons Data Portal at https://portal.gdc.cancer.gov. Datasets CRC-FFPE-FIN and CRCOYS-FIN are not publicly available due to institutional ethical restrictions protecting patient privacy. The code used in this study is publicly available at: https://github.com/Keski-Suomen-AI-Hub-II/digital-pathology-CRC/tree/main/multi-scale-MMR.

Ethics statement

This study did not involve any experiments on human participants or animals, and no ethical approval was required. All data used were fully anonymized and obtained from publicly available sources or institutional repositories under appropriate data use agreements. No identifiable personal or clinical information was accessed or used in this research.

CRediT authorship contribution statement

Liisa Petäinen: Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Juha P. Väyrynen:** Writing – review & editing, Supervision, Resources, Investigation, Data curation, Conceptualization. **Jan Böhm:** Writing – review & editing, Supervision, Resources, Investigation, Data curation, Conceptualization. **Pekka Ruusuvuori:** Writing – review & editing, Methodology, Formal analysis, Conceptualization. **Maarit Ahtiainen:** Writing – review & editing, Resources, Investigation, Data curation. **Hanna Elomaa:** Writing – review & editing, Resources. **Henna Karjalainen:** Writing – review & editing, Resources. **Meeri Kastinen:** Writing – review & editing, Resources. **Vilja V. Tapiainen:** Writing – review & editing, Resources. **Ville K. Äijälä:** Writing – review & editing, Resources. **Päivi Sirniö:** Writing – review & editing, Resources. **Anne Tuomisto:** Writing – review & editing, Resources.

Markus J. Mäkinen: Writing – review & editing, Resources, Conceptualization. **Jukka-Pekka Mecklin:** Writing – review & editing, Resources, Conceptualization. **Ilkka Pölönen:** Writing – review & editing, Supervision, Methodology, Formal analysis, Conceptualization. **Sami Äyrämö:** Writing – review & editing, Supervision, Resources, Project administration, Funding acquisition, Formal analysis, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

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

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.cmpb.2026.109317.

References

[1] V. Baxi, R. Edwards, M. Montalto, et al., Digital pathology and artificial intelligence in translational medicine and clinical practice, Mod. Pathol. 35 (1) (2022) 23–32.
[2] O. Koteluk, A. Wartecki, S. Mazurek, et al., How do machines learn? Artificial intelligence as a new era in medicine, J. Pers. Med. 11 (1) (2021) 32.
[3] C.L. Srinidhi, O. Ciga, A.L. Martel, Deep neural network models for computational histopathology: a survey, Med. Image Anal. 67 (2021) 101813.
[4] P.C. Rizzo, A. Caputo, E. Maddalena, et al., Digital pathology world tour, Digit. Health 9 (2023) 20552076231194551.
[5] C. McGenity, E.L. Clarke, C. Jennings, et al., Artificial intelligence in digital pathology: a systematic review and meta-analysis of diagnostic test accuracy, npj Digit. Med. 7 (1) (2024) 114.
[6] H. Kawakami, A. Zaanan, F.A. Sinicrope, Microsatellite instability testing and its role in the management of colorectal cancer, Curr. Treat. Options Oncol. 16 (2015) 1–15.
[7] G. Argiles, J. Tabernero, R. Labianca, et al., Localised colon cancer: esmo clinical practice guidelines for diagnosis, treatment and follow-up, Ann. Oncol. 31 (10) (2020) 1291–1305.
[8] A. Echle, N.G. Laleh, P.L. Schrammen, et al., Deep learning for the detection of microsatellite instability from histology images in colorectal cancer: a systematic literature review, ImmunoInformatics 3 (2021) 100008.
[9] A.J. Kacew, G.W. Strohbehn, L. Saulsberry, et al., Artificial intelligence can cut costs while maintaining accuracy in colorectal cancer genotyping, Front. Oncol. 11 (2021) 630953.
[10] R. Cao, F. Yang, S.C. Ma, et al., Development and interpretation of a pathomics-based model for the prediction of microsatellite instability in colorectal cancer, Theranostics 10 (24) (2020) 11080.
[11] A. Echle, H.I. Grabsch, P. Quirke, et al., Clinical-grade detection of microsatellite instability in colorectal tumors by deep learning, Gastroenterology 159 (4) (2020) 1406–1416.
[12] J.N. Kather, A.T. Pearson, N. Halama, et al., Deep learning can predict microsatellite instability directly from histology in gastrointestinal cancer, Nat. Med. 25 (7) (2019) 1054–1056.
[13] S.H. Lee, I.H. Song, H.J. Jang, Feasibility of deep learning-based fully auto- mated classification of microsatellite instability in tissue slides of colorectal cancer, Int. J. Cancer 149 (3) (2021) 728–740.
[14] R. Yamashita, J. Long, T. Longacre, et al., Deep learning model for the prediction of microsatellite instability in colorectal cancer: a diagnostic study, Lancet Oncol. 22 (1) (2021) 132–141.
[15] K. He, X. Zhang, S. Ren, et al., Deep residual learning for image recognition, in: Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2016, pp. 770–778.
[16] N. Hashimoto, D. Fukushima, R. Koga, et al., Multi-scale domain-adversarial multiple-instance cnn for cancer subtype classification with unannotated histopatho- logical images, in: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2020, pp. 3852–3861.
[17] N. Marini, S. Otálora, F. Ciompi, et al., Multi-scale task multiple instance learning for the classification of digital pathology images with global annotations, in: Proc. MICCAI Workshop Comput. Pathol., 2021, pp. 170–181. PMLR.
[18] M. Valkonen, K. Kartasalo, K. Liimatainen, et al., Dual structured convolutional neural network with feature augmentation for quantitative characterization of tissue histology, in: Proc. IEEE Int. Conf. Comput. Vis. Workshops, 2017, pp. 27–35.
[19] F. Gu, N. Burlutskiy, M. Andersson, and L.K. Wilen, Multi-resolution networks for semantic segmentation in whole slide images, Proc. Comput. Pathol. Ophthalmic Med. Image Anal.: First Int. Workshop COMPAY 2018 5th Int. Workshop, 5, 2018, pp. 11–18 OMIA 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16-20, 2018, ProceedingsSpringer.
[20] R. Schmitz, F. Madesta, M. Nielsen, et al., Multi-scale fully convolutional neural networks for histopathology image segmentation: from nuclear aberrations to the global tissue architecture, Med. Image Anal. 70 (2021) 101996.
[21] H. Tokunaga, Y. Teramoto, A. Yoshizawa, et al., Adaptive weighting multi-field-of-view cnn for semantic segmentation in pathology, in: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2019, pp. 12597–12606.
[22] M. Van Rijthoven, M. Balkenhol, K. Silin¸a, et al., Hooknet: multi-resolution convolutional neural networks for semantic segmentation in histopathology whole-slide images, Med. Image Anal. 68 (2021) 101890.
[23] W.S. Shim, K. Yim, T.J. Kim, et al., Deeprepath: identifying the prognostic features of early-stage lung adenocarcinoma using multi-scale pathology images and deep convolutional neural networks, Cancers 13 (13) (2021) 3308.
[24] S.C. Kosaraju, J. Hao, H.M. Koh, M. Khang, Deep-hipo: multi-scale receptive field deep learning for histopathological image analysis, Methods 179 (2020) 3–13.
[25] T.S. Sheikh, Y. Lee, M. Cho, Histopathological classification of breast cancer images using a multi-scale input and multi-feature network, Cancers 12 (8) (2020) 2031.
[26] M.S. Jain, T.F. Massoud, Predicting tumour mutational burden from histopathological images using multiscale deep learning, Nat. Mach. Intell. 2 (6) (2020) 356–362.
[27] R. Cao, Q. Gu, D. Tan, et al., Prediction of microsatellite instability of colorectal cancer using multi-scale pathological images based on deep learning, in: Proc. 2022 IEEE Int. Conf. Bioinform. Biomed. (BIBM), 2022, pp. 1461–1466. IEEE.
[28] M. Bilal, S.E.A. Raza, A. Azam, et al., Development and validation of a weakly supervised deep learning framework to predict the status of molecular pathways and key mutations in colorectal cancer from routine histology images: a retrospective study, Lancet Digit. Health 3 (12) (2021) e763–e772.
[29] M.A. Jenkins, S. Hayashi, O'shea AM, et al., Pathology features in bethesda guidelines predict colorectal cancer microsatellite instability: a population-based study, Gastroenterology 133 (1) (2007) 48–56.
[30] H. Elomaa, M. Ahtiainen, S.A. Väyrynen, et al., Prognostic significance of spatial and density analysis of t lymphocytes in colorectal cancer, Br. J. Cancer 127 (3) (2022) 514–523.
[31] J. Liu, T. Lichtenberg, K.A. Hoadley, et al., An integrated tcga pan-cancer clinical data resource to drive high-quality survival outcome analytics, Cell 173 (2) (2018) 400–416.
[32] L. Petäinen, J.P. Väyrynen, P. Ruusuvuori, et al., Domain-specific transfer learning in the automated scoring of tumor-stroma ratio from histopathological images of colorectal cancer, Plos one 18 (5) (2023) e0286270.
[33] M. Macenko, M. Niethammer, J.S. Marron, et al., A method for normalizing histology slides for quantitative analysis, in: Proc. 2009 IEEE Int. Symp. Biomed. Imaging: Nano Macro, 2009, pp. 1107–1110. IEEE.
[34] A. Howard, M. Sandler, G. Chu, et al., Searching for mobilenetv3, in: Proc. IEEE/CVF Int. Conf. Comput. Vis., 2019, pp. 1314–1324.
[35] R.J. Chen, T. Ding, M.Y. Lu, et al., Towards a general-purpose foundation model for computational pathology, Nat. Med. 30 (3) (2024) 850–862.
[36] C. Saillard, R. Jenatton, L. Llinares, F.·opez, et al., H-optimus-0, 2024, URL: https://github.com/bioptimus/releases/tree/main/models/h-optimus/v0. (accessed November 15, 2025).
[37] O. Russakovsky, J. Deng, H. Su, et al., Imagenet large scale visual recognition challenge, Int. J. Comput. Vis. 115 (2015) 211–252.
[38] M. Ilse, J. Tomczak, M. Welling, Attention-based deep multiple instance learning, in: Proc. Int. Conf. Mach. Learn., 2018, pp. 2127–2136. PMLR.
[39] M.Y. Lu, D.F. Williamson, T.Y. Chen, et al., Data-efficient and weakly supervised computational pathology on whole-slide images, Nat. Biomed. Eng. 5 (6) (2021) 555–570.
[40] H. Li, J. Qin, Z. Li, et al., Systematic review and meta-analysis of deep learning for MSI-H in colorectal cancer whole slide images, npj Digit. Med. 8 (1) (2025) 456.
[41] H. Wei, X. Zhang, Z. Zhou, et al., Hybrid model for predicting microsatellite instability in colorectal cancer using hematoxylin & eosin-stained images and clinical features, Front. Oncol. 15 (2025) 1580195.