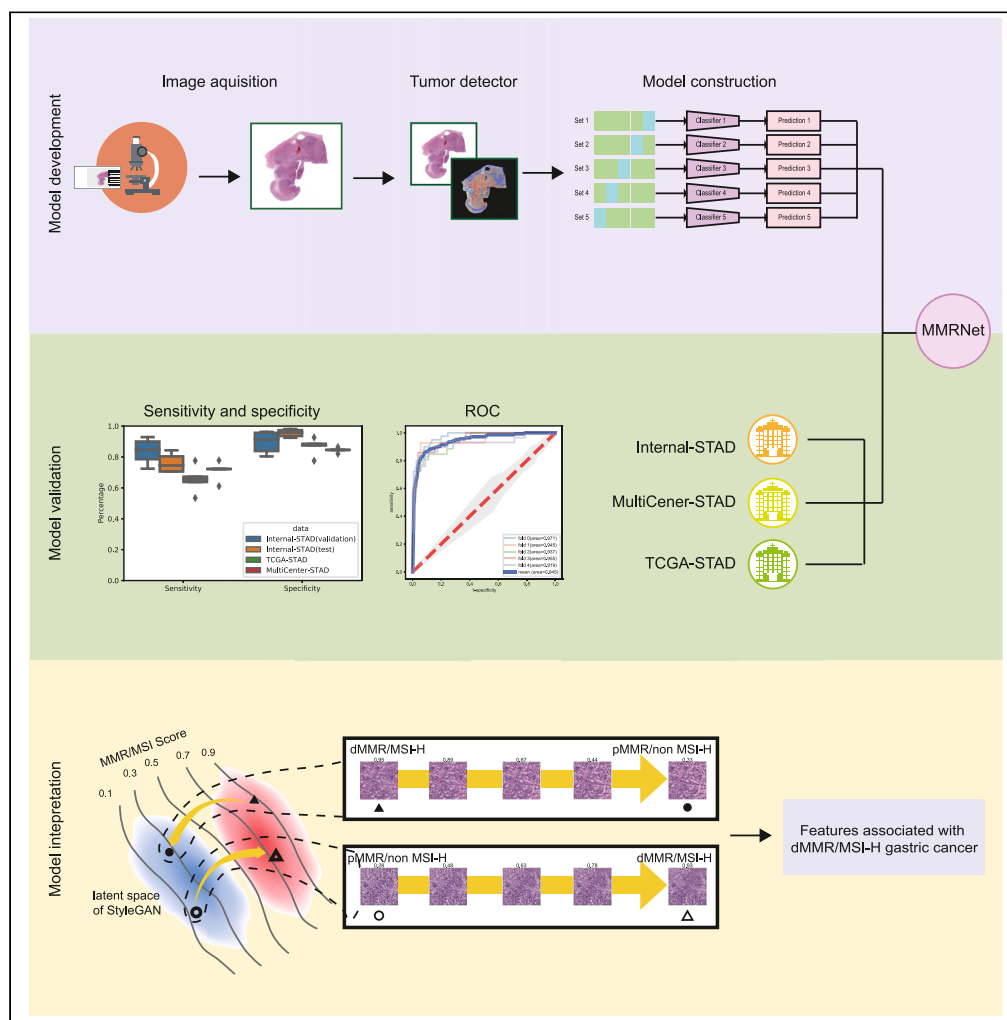## Article

An interpretable deep learning model for identifying the morphological characteristics of dMMR/MSI-H gastric cancer

Xueyi Zheng, Bingzhong Jing, Zihan Zhao, ..., Chaofeng Li, Dan Xie, Muyan Cai

lichaofeng@sysucc.org.cn (C.L.)
xiedan@sysucc.org.cn (D.X.)
caimy@sysucc.org.cn (M.C.)

Highlights

MMRNet can automatically predict MMR/MSI status from gastric cancer H&E-stained WSIs

A mapping approach MMRMapping can efficiently interpret MMRNet

The interpretable method can aid the growth of young pathologists

Article

An interpretable deep learning model for identifying the morphological characteristics of dMMR/MSI-H gastric cancer

Xueyi Zheng,[1,13] Bingzhong Jing,[2,13] Zihan Zhao,[1,13] Ruixuan Wang,[3,13] Xinke Zhang,[1] Haohua Chen,[2] Shuyang Wu,[1] Yan Sun,[4] Jiangyu Zhang,[5] Hongmei Wu,[6] Dan Huang,[7] Wenbiao Zhu,[8] Jianning Chen,[9] Qinghua Cao,[10] Hong Zeng,[11] Jinling Duan,[1] Yuanliang Luo,[1] Zhicheng Li,[1] Wuhao Lin,[1] Runcong Nie,[12] Yishu Deng,[2] Jingping Yun,[1] Chaofeng Li,[2,*] Dan Xie,[1,*] and Muyan Cai[1,14,*]

SUMMARY

Accurate tumor diagnosis by pathologists relies on identifying specific morphological characteristics. However, summarizing these unique morphological features in tumor classifications can be challenging. Although deep learning models have been extensively studied for tumor classification, their indirect and subjective interpretation obstructs pathologists from comprehending the model and discerning the morphological features accountable for classifications. In this study, we introduce a new approach utilizing Style Generative Adversarial Networks, which enables a direct interpretation of deep learning models to detect significant morphological characteristics within datasets representing patients with deficient mismatch repair/microsatellite instability-high gastric cancer. Our approach effectively identifies distinct morphological features crucial for tumor classification, offering valuable insights for pathologists to enhance diagnostic accuracy and foster professional growth.

INTRODUCTION

Diagnostic pathology is universally regarded as the gold standard for tumor identification, and its efficacy relies heavily on the ability of pathologists to recognize the morphological characteristics of various tumor subtypes and make accurate diagnoses.[1] However, pathologists, particularly those in small centers, may have limited opportunities to encounter diverse tumor classifications, and even experienced pathologists face the challenge of summarizing specific morphological characteristics due to the abundance of information contained within pathological slides. As a result, the development of an effective tool that can aid pathologists in identifying the histopathological characteristics of tumor classifications and help them gain the necessary expertise is of paramount importance.

Convolutional neural networks have shown great potential as a promising tool for extracting relevant features from histology slides, which can be used to accurately classify tumors with high precision.[2,3] Despite the promising diagnostic performances of deep learning models in tumor diagnosis,[4,5] their black-box nature often limits the comprehensibility of the features extracted by the models. This presents a challenge for pathologists who require clear and comprehensive features to understand the reasoning behind the model's decision.[6,7] The lack of clear interpretation raises important questions about the effectiveness of deep learning models in aiding pathologists to recognize the morphological characteristics of various tumor subtypes.[4,8] To address this issue, it is critical to develop new methods that can help interpret deep learning models and extract meaningful features to facilitate the growth of pathologists.

[1]Department of Pathology, State Key Laboratory of Oncology in South China, Guangdong Provincial Clinical Research Center for Cancer, Sun Yat-sen University Cancer Center, Guangzhou 510060, China
[2]Artificial Intelligence Laboratory, State Key Laboratory of Oncology in South China, Guangdong Provincial Clinical Research Center for Cancer, Sun Yat-sen University Cancer Center, Guangzhou 510060, China
[3]School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou 510006, China
[4]Department of Pathology, Tianjin Medical University Cancer Institute and Hospital, Tianjin 300000, China
[5]Department of Pathology, Affiliated Cancer Hospital & Institute of Guangzhou Medical University, Guangzhou 510095, China
[6]Department of Pathology, Guangdong Provincial People's Hospital, Guangdong Academy of Medical Sciences, Guangzhou 510080, China
[7]Department of Pathology, Fudan University Shanghai Cancer Center, Shanghai 200032, China
[8]Department of Pathology, Shantou University medical college Meizhou clinical school, Meizhou People's Hospital, Meizhou 514011, China
[9]Department of Pathology, The Third Affiliated Hospital, Sun Yat-sen University, Guangzhou 510635, China
[10]Department of Pathology, The First Affiliated Hospital, Sun Yat-sen University, Guangzhou 510080, China
[11]Department of Pathology, Sun Yat-Sen Memorial Hospital, Sun Yat-Sen University, Guangzhou 510120, China
[12]Department of Department of Gastric Surgery, State Key Laboratory of Oncology in South China, Guangdong Provincial Clinical Research Center for Cancer, Sun Yat-sen University Cancer Center, Guangzhou 510060, China
[13]These authors contributed equally
[14]Lead contact
*Correspondence: lichaofeng@sysucc.org.cn (C.L.), xiedan@sysucc.org.cn (D.X.), caimy@sysucc.org.cn (M.C.)
https://doi.org/10.1016/j.isci.2024.109243

Table 1. Diagnostic performance of MMRNet predicting MMR/MSI status in gastric cancer cohorts

Cohorts	AUROC	Sensitivity	Specificity	NPV	PPV
Internal-STAD	0.930 (0.923, 0.938)	0.761 (0.685, 0.837)	0.955 (0.925, 0.985)	0.954 (0.942, 0.967)	0.776 (0.671, 0.881)
MutiCenter-STAD	0.895 (0.874, 0.917)	0.711 (0.636, 0.787)	0.844 (0.825, 0.864)	0.963 (0.955, 0.972)	0.337 (0.307, 0.367)
TCGA-STAD	0.844 (0.830, 0.858)	0.655 (0.548, 0.762)	0.868 (0.799, 0.937)	0.905 (0.885, 0.926)	0.582 (0.501, 0.663)

95% confidence intervals were included in brackets. MMR, mismatch repair protein; MSI, microsatellite instability; Internal-STAD, an internal cohort from a single medical center; MultiCenter-STAD, an external cohort from multiple medical centers; TCGA-STAD, an external cohort from The Cancer Genome Atlas. AUROC, area under the receiver operating curve; NPV, negative predictive value; PPV, positive predictive value.

Generative adversarial network (GAN) model is currently proposed for generating realistic images. GAN consists of two main components: the Generator and the Discriminator.[9] The Generator is responsible for creating synthetic images similar to real images by learning statistical features and structures from the dataset, while the Discriminator acts as a binary classifier to evaluate the authenticity of images. Through learning features from images, the GAN obtains the latent space,[10] which shows an interpretable structure and allows semantic vector operations that translate into tissue feature transformations. This capability of GAN finds extensive applications in model interpretation, which may help visualize how the deep learning model make decisions.

In this study, we have developed a framework to visualize the dynamic diagnostic process of the deep learning model and capture histopathological features of tumor classifications, thus aiding the growth of pathologists. Our approach involved building a deep learning model MMRNet to extract patch-level and then slide-level diagnosis of deficient mismatch repair (dMMR)/microsatellite instability-high (MSI-H) gastric cancer (GC), a distinct GC molecular subtype which can benefit from immunotherapy. The model performances were further validated on whole slide images (WSIs) from two external cohorts. To reveal the decision-making progress of MMRNet, we introduced a model interpretation method. First, we trained a Style Generative Adversarial Network (StyleGAN)[11] model to generate realistic images capturing the rich tissue structures of dMMR/MSI-H GC. Then, using the MMRMapping conditional regression-based mapping model, we manipulated the morphological transformation from dMMR/MSI-H GC images to proficient mismatch repair (pMMR)/non MSI-H GC images in the latent space of StyleGAN, explicitly displaying the morphological features responsible for MMRNet's classifications. To localize significant histopathological characteristics related to the classification, we developed the classification activation map (CAM)-blending algorithm in the inference stage, which calculated the CAM of MMRNet[12] to guide the local blending of original image and the transformed image. The CAM-blending algorithm facilitated the transformation in the high-attention areas while inhibiting transformation in the low-attention areas, thus helping pathologists review and recognize significant morphological characteristics of dMMR/MSI-H GC. Finally, we further evaluated the diagnostic performances of junior pathologists in predicting dMMR/MSI-H GC before and after learning the morphological characteristics.

## RESULTS

Study participants

Our method for identifying morphological features of dMMR/MSI-H GC was designed to be broadly applicable across various tumor classifications. The Internal-STAD dataset comprised 202 hematoxylin & eosin (H&E)-stained WSIs from 105 patients diagnosed with dMMR/MSI-H GC and 1060 WSIs from 562 patients with pMMR/non-MSI-H GC. In the MultiCenter-STAD dataset, 180 WSIs were gathered from 180 patients, including 18 patients with dMMR/MSI-H GC and 162 patients with pMMR/non-MSI-H GC. Conversely, the TCGA-STAD dataset encompassed 284 WSIs obtained from 284 GC patients, with 60 patients diagnosed with dMMR/MSI-H GC and 224 patients with pMMR/non-MSI-H GC.

Diagnostic performances of MMRNet

Given the morphological features like lymphocytic aggregates might be indicative of the tumor response contributing to dMMR/MSI-H subtype, as previously descripted in relevant studies,[4,8] we did not present objective metrics for the tumor tissue classifier. To predict MMR/MSI status from GC WSIs, we built MMRNet using ResNet18[13] for training and internal validation. The 5-fold cross-validation on each fold produced an AUROC range from 0.919 to 0.971 (Table S1) on Internal-STAD. Upon all test folds, the MMRNet obtained an averaged AUROC of 0.930 [95% confidence interval (CI): 0.923–0.938], sensitivity of 0.761 (95%CI: 0.685–0.837), specificity of 0.955 (95%CI: 0.925–0.985), positive predictive value (PPV) of 0.776 (95% CI: 0.671–0.881) and negative predictive value (NPV) of 0.954 (95%CI: 0.942–0.967). On MultiCenter-STAD, MMRNet achieved an averaged AUROC of 0.895 (95% CI: 0.874–0.917), sensitivity of 0.711 (95%CI: 0.636–0.787), specificity of 0.844 (95%CI: 0.825–0.864), PPV of 0.337 (95%CI: 0.307–0.367), and NPV of 0.963 (95%CI: 0.955, 0.972). On TCGA-STAD, MMRNet yielded an averaged AUROC of 0.844 (95%CI: 0.830–0.858), sensitivity of 0.655 (95%CI: 0.548–0.762), specificity of 0.868 (95%CI: 0.799–0.937), PPV of 0.582 (95% CI: 0.501–0.663), and NPV of 0.905 (95%CI: 0.885, 0.926) (Table 1; Figures 1 and S1).

Model interpretation summarizing histopathological characteristics of dMMR/MSI-H GC

To summarize histopathological characteristics of dMMR/MSI-H GC, we employed MMRMapping to transform dMMR/MSI-H images to pMMR/non MSI-H images to output the manipulation direction vector of the input image generated by StyleGAN. CAM-blending was utilized
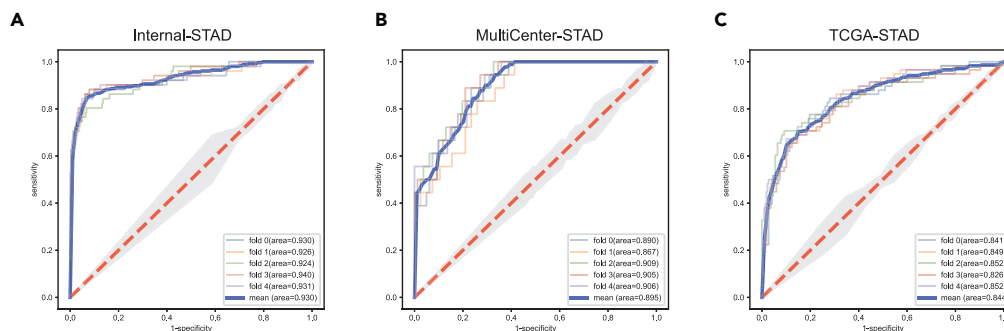
Figure 1. Performance of MMRNet on internal and external testing cohorts

(A) On internal test set of Internal-STAD dataset, the MMRNet achieved an averaged AUROC of 0.930 (95%CI: 0.923–0.938).

(B) On MultiCenter-STAD dataset, the MMRNet achieved an averaged AUROC of 0.895 (95%CI: 0.874–0.917).

(C) On TCGA-STAD dataset, the MMRNet achieved an averaged AUROC of 0.844 (95%CI: 0.830–0.858). MMR, mismatch repair; Internal-STAD, an internal cohort from a single medical center; MultiCenter-STAD, an external cohort from multiple medical centers; TCGA-STAD, an external cohort from The Cancer Genome Atlas; CI, confidence interval.

to localize the morphological changes in the high-attention areas of the heatmap, where pathologists can recognize the morphological features that were closely related to the results of MMRNet. After reviewing 10,000 patch groups generated by StyleGAN, two expert pathologists identified seven features associated with dMMR/MSI-H GC, including syncytial cells, tumor infiltrative lymphocytes, lymphoid stroma, medullary histology, vacuolar nucleus, recognizable nucleolus and mucinous differentiation (Figures 2 and S2–S8).

To investigate the correlation between the previously mentioned morphological features and dMMR/MSI-H GC, two senior pathologists collaborated to determine the presence of each feature in WSIs of two external cohorts. Results from the MultiCenter-STAD cohort showed that dMMR/MSI-H GCs were more likely to presence of syncytial cells (p = 0.023), tumor infiltrative lymphocytes (p = 0.026), vacuolar nucleus (p = 0.004), and recognizable nucleolus (p = 0.001) compared to pMMR/non MSI-H GCs. Similar findings were observed on the TCGA-STAD cohort, with dMMR/MSI-H GCs possessing the same significant features, including syncytial cells (p = 0.006), tumor infiltrative lymphocytes (p < 0.001), vacuolar nucleus (p < 0.001), and recognizable nucleolus (p < 0.001) (Figure 3; Table 2). These data suggest that the interpretable method used to summarize the morphological features is reliable and can be used to identify dMMR/MSI-H GCs.
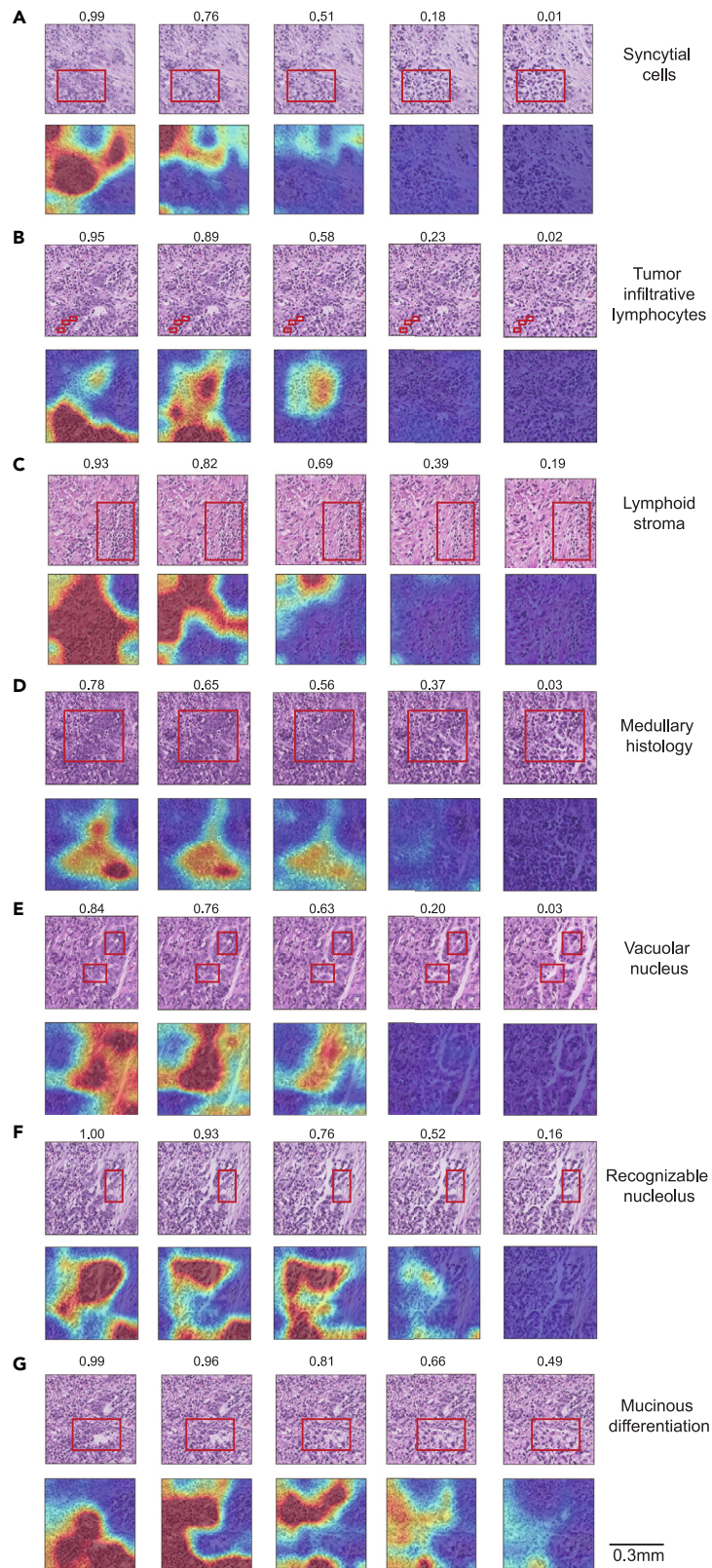
Model interpretation aiding the growth of young pathologists

To assess the effectiveness of the interpretable method in aiding the growth of young pathologists, a reader study was conducted. A test set of 60 WSIs was reviewed by four junior pathologists who were asked to classify each GC image as either dMMR/MSI-H GC or pMMR/non MSI-H GC based on their experience. They were then provided with a visual presentation illustrating distinct summarized morphological features and asked if they wanted to change their initial assessment. Before learning morphological features, Junior Pathologist 1 achieved an AUROC of 0.517 (95%CI: 0.384–0.648), which improved significantly to an AUROC of 0.683 (95%CI: 0.550–0.797) after learning morphological features (p = 0.004). Junior Pathologist 2 yielded AUROCs of 0.567 (95%CI: 0.432–0.694) and 0.650 (95%CI: 0.516–0.769) before and after learning the morphological features, respectively. Junior Pathologist 3 obtained AUROCs of 0.550 (95%CI: 0.416–0.679) and 0.733 (95%CI: 0.603–0.839) before and after learning the morphological features, respectively, with a significant improvement (p = 0.042). Junior Pathologist 4 also showed a significant improvement with an AUROC of 0.667 (95%CI: 0.533–0.783) after learning, compared to an AUROC of 0.550 (95% CI: 0.416–0.679) before learning (p = 0.016). The diagnostic performances of the junior pathologists were summarized in Table S2, indicating that the interpretable method of the deep learning model can be a valuable tool in improving the performance of young pathologists.

DISCUSSION

We introduce a new method for interpreting deep learning models that allows pathologists to identify morphological characteristics of dMMR/MSI-H GC in an understandable and interpretable manner. This report marks the initial examination of pathologists' abilities in identifying MMR/MSI status through H&E-stained slides. While deep learning models have shown superior performance in tumor classifications, surpassing even human experts,[14,15] their lack of interpretability has hindered their clinical application.[2] The ability to identify the properties that contribute to the model's predictions can help advance the understanding of underlying biological processes. To address this issue, we propose a new mapping approach based on StyleGAN-generated images to interpret the deep learning model. GANs have achieved state-of-the-art performance in various image processing and analysis tasks.[9] We leverage this technology for model interpretability by designing an offset vector for the target latent code, which visualizes how the deep learning model makes decisions. This is the first attempt to apply a mapping method in GANs, which provides an explainable way to highlight significant morphological features of the predictive targets.

Identifying morphological characteristics of different tumor subtypes is a critical aspect of accurate pathological diagnosis. It is indeed pertinent to acknowledge that while some pathologists might possess inherent abilities to distinguish between dMMR/MSI-H GC based on their expertise and prior knowledge, this capability might not be universally consistent across all pathologists or diagnostic settings.

A. Syncytial cells — 0.99, 0.76, 0.51, 0.18, 0.01
B. Tumor infiltrative lymphocytes — 0.95, 0.89, 0.58, 0.23, 0.02
C. Lymphoid stroma — 0.93, 0.82, 0.69, 0.39, 0.19
D. Medullary histology — 0.78, 0.65, 0.56, 0.37, 0.03
E. Vacuolar nucleus — 0.84, 0.76, 0.63, 0.20, 0.03
F. Recognizable nucleolus — 1.00, 0.93, 0.76, 0.52, 0.16
G. Mucinous differentiation — 0.99, 0.96, 0.81, 0.66, 0.49

0.3mm

Figure 2. The morphological features related to dMMR/MSI-H gastric cancer detected by synthetic patch groups

(A–G) Each patch group contained five H&E-stained patches (upper panels) and their corresponding heatmaps (lower panels), with a scale bar of 0.3mm. By manipulating morphological changes in the high-attention areas of heatmap (indicated by the red boxes in H&E-stained patches), MMRMapping gradually decreased the predictive scores of the H&E-stained patches from dMMR/MSI-H scores to pMMR/non MSI-H scores. Pathologists reviewed these patch groups and summarized seven features associated with dMMR/MSI-H gastric cancer, including syncytial cells, tumor infiltrative lymphocytes, lymphoid stroma, medullary histology, vacuolar nucleus, recognizable nucleolus, and mucinous differentiation, labeled from A to G. H&E, hematoxylin & eosin; dMMR, deficient mismatch repair; MSI-H, microsatellite instability-high; pMMR, proficient mismatch repair.

Therefore, the summarization of features should consider the diverse expertise levels and variations in the interpretative abilities of pathologists, ensuring that the results and conclusions are not solely dependent on this variable factor. Our method is not limited to dMMR/MSI-H GC but can be applied to a broad range of tumor subtypes. By learning the diagnostic mechanism of deep learning models and capturing significant morphological features of different tumor subtypes with the aid of our interpretation method, junior pathologists can improve the efficiency and accuracy of their diagnoses.

3mm   pMMR/non MSI-H   dMMR/MSI-H   50um
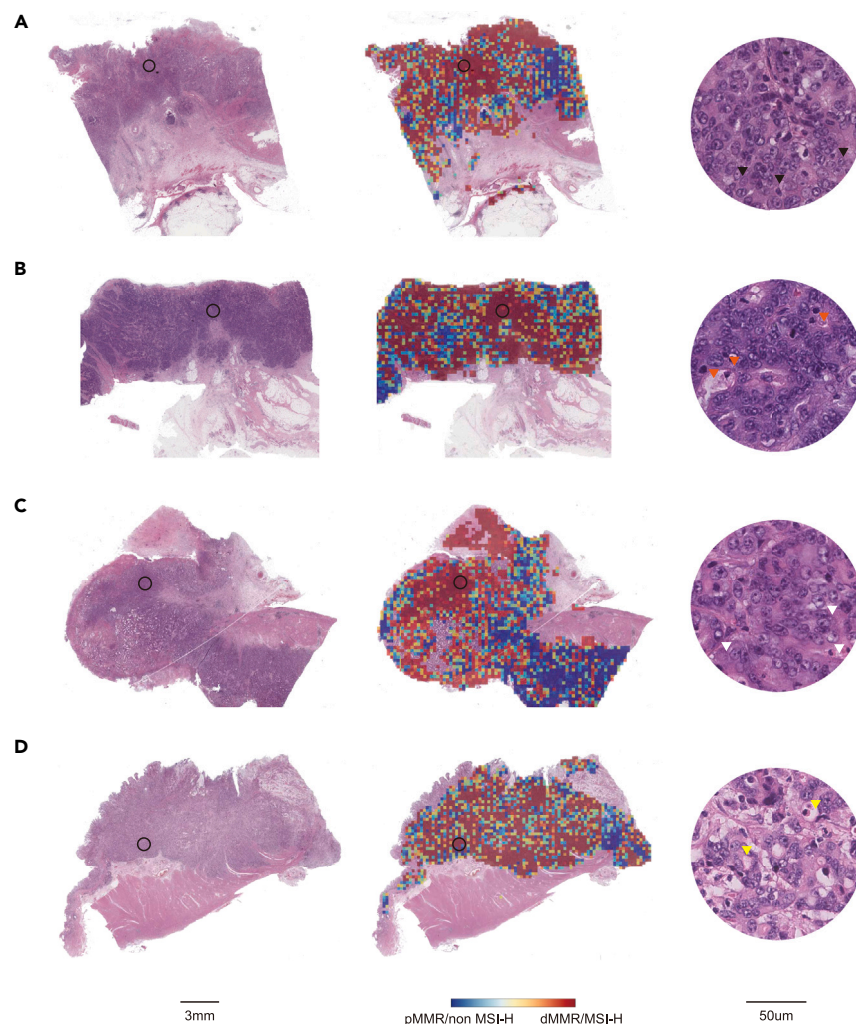
Figure 3. The morphological features related to dMMR/MSI-H gastric cancer in whole slide images

(A–D) The left panels of figure displayed representative H&E-stained WSIs (scale bar, 3mm) of patients with dMMR/MSI-H gastric cancer. The heatmaps overlapped on these WSIs (middle panels) showed that tumor tiles were mainly predicted as dMMR/MSI-H gastric cancer with a high score (red color). Tiles with a high score (black circle corresponding in H&E-stained WSIs) were mainly focused on the areas (right panels; scale bar, 50um) of syncytial cells (black arrowhead), tumor infiltrative lymphocytes (orange arrowhead), vacuolar nucleus (white arrowhead) and recognizable nucleolus (yellow arrowhead), from A to D. H&E, hematoxylin & eosin; WSI, whole slide images; dMMR, deficient mismatch repair; MSI-H, microsatellite instability-high; pMMR, proficient mismatch repair.

Table 2. Validation of the morphological features associated with dMMR/MSI-H gastric cancer on two external cohorts

Morphological features		MultiCenter-STAD			TCGA-STAD		
		pMMR/non MSI-H (n = 162)	dMMR/MSI-H (n = 18)	*p* value	pMMR/non MSI-H (n = 218)	dMMR/MSI-H (n = 58)	*p* value
Syncytial cells	Absence	143 (88.3%)	12 (66.7%)	0.023*	198 (90.8%)	44 (75.9%)	0.006*
	Presence	19 (11.7%)	6 (33.3%)		20 (9.2%)	14 (24.1%)	
Tumor infiltrative lymphocytes	Absence	91 (56.2%)	5 (27.8%)	0.026*	154 (70.6%)	5 (8.6%)	<0.001*
	Presence	71 (43.8%)	13 (72.2%)		64 (29.4%)	53 (91.4%)	
Lymphoid stroma	Absence	134 (82.7%)	17 (94.4%)	0.314	209 (95.9%)	58 (100%)	0.212
	Presence	28 (17.3%)	1 (5.6%)		9 (4.1%)	0 (0%)	
Medullary histology	Absence	149 (92%)	17 (94.4%)	1.000	213 (97.7%)	58 (100%)	0.587
	Presence	13 (8%)	1 (5.6%)		5 (2.3%)	0 (0%)	
Vacuolar nucleus	Absence	49 (30.2%)	0 (0%)	0.004*	86 (39.4%)	2 (3.4%)	<0.001*
	Presence	113 (69.8%)	18 (100%)		132 (60.6%)	56 (96.6%)	
Recognizable nucleolus	Absence	56 (34.6%)	0 (0%)	0.001*	102 (46.8%)	3 (5.2%)	<0.001*
	Presence	106 (65.4%)	18 (100%)		116 (53.2%)	55 (94.8%)	
Mucinous differentiation	Absence	137 (84.6%)	17 (94.4%)	0.478	180 (82.6%)	43 (74.1%)	0.188
	Presence	25 (15.4%)	1 (5.6%)		38 (17.4%)	15 (25.9%)	

Percentage was included in brackets. MultiCenter-STAD, an external cohort from multiple medical centers; TCGA-STAD, an external cohort from The Cancer Genome Atlas; dMMR, deficient mismatch repair; MSI-H, microsatellite instability-high; pMMR, proficient mismatch repair. *p < 0.05 was considered significant difference.

Limitations of the study

While our study demonstrates promising results, we also acknowledge some limitations that need to be addressed. First, relative to magnifications of 20× or 40×, employing a 10× magnification level in image patches tends to capture larger-scale visual features, potentially leading to the loss or reduction of finer details within these patches.[14] Second, while MMRNet has delivered satisfactory results, we acknowledged the importance of tailoring the threshold for specific clinical purposes, including attainment of high sensitivity to allow reliable identification of patients who definitely should not undergo further MMR/MSI testing. This personalized threshold adjustment could enhance the clinical utility of the algorithm, especially in distinguishing between different molecular subtypes of GC. Third, we used synthetic patches to summarize morphological features instead of real patches. While synthetic patches have been shown to be useful in complementing real data, subtle differences between synthetic and real patches exist. However, the fidelity of synthetic patches generated by StyleGAN has been well studied and are commonly used as a complement of real data with more manipulatable properties.[16,17] Fourth, the modified offset vector's influence on multiple features simultaneously posed challenges in isolating them individually. To navigate this, pathologists collected all discernible features that gradually faded from the patch group. Fifth, we did not conduct further validation on the accuracy of the tumor detection process since previous study has demonstrated its effectiveness in distinguishing between tumor and normal tissue using deep learning models, which might potentially influence the results of MMRNet. Lastly, the morphological features identified in our study were limited to those recognizable by the naked eye, and there may be more subtle features that the deep learning model uses, which could result in the performance differences between the model and young pathologists. Incorporating more objective methods in the future involves modifying just one histopathological characteristic within a specific patch group or automating feature aggregation from generated patches. Although there is still a long way in terms of model interpretability, our study represents a promising attempt to address this issue and lays foundation for future research.

In summary, we have presented a new method for interpreting deep learning models, which can aid pathologists in efficiently capturing specific morphological characteristics and thus facilitate their growth.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability

- EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS
  ○ Human subjects
- METHOD DETAILS
  ○ Slide scanning
  ○ Determination of MMR/MSI status
  ○ Tumor detector
  ○ MMRNet development and validation
  ○ Model interpretability
  ○ Identification and validation of the morphological features related to dMMR/MSI-H GC
  ○ Reader study
- QUANTIFICATION AND STATISTICAL ANALYSIS

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2024.109243.

AUTHOR CONTRIBUTIONS

M.Y.C., D.X., X.Y.Z., and C.F.L. conceptualized and designed the study. B.Z.J., Z.H.Z., R.X.W., H.H.C., S.Y.W., and Y.S.D. developed and tested the deep learning model. X.Y.Z., X.K.Z., and R.C.N. analyzed the data. Y.S., J.Y.Z., H.M.W., D.H., W.B.Z., J.N.C., Q.H.C., H.Z., J.L.D., Y.L.L., Z.C.L., W.H.L., and J.P.Y. performed the investigation, validation, formal analysis, and visualization. X.Y.Z. and B.Z.J. drafted the manuscript, which was edited and critically reviewed by M.Y.C., C.F.L. D.X., Z.H.Z., R.X.W., X.K.Z., H.H.C., and S.Y.W. All authors read and approved the final manuscript and had final responsibility for the decision to submit for publication.

DECLARATION OF INTERESTS

All authors declare no financial or non-financial competing interests.

REFERENCES

1. Zhang, Z., Chen, P., Mcgough, M., Xing, F., Wang, C., Bui, M., Xie, Y., Sapkota, M., Cui, L., Dhillon, J., et al. (2019). Pathologist-level interpretable whole-slide cancer diagnosis with deep learning. Nat. Mach. Intell. *1*, 236–245.

2. Bera, K., Schalper, K.A., Rimm, D.L., Velcheti, V., and Madabhushi, A. (2019). Artificial intelligence in digital pathology - new tools for diagnosis and precision oncology. Nat. Rev. Clin. Oncol. *16*, 703–715.

3. LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. Nature *521*, 436–444. https://doi.org/10.1038/nature14539.

4. Yamashita, R., Long, J., Longacre, T., Peng, L., Berry, G., Martin, B., Higgins, J., Rubin, D.L., and Shen, J. (2021). Deep learning model for the prediction of microsatellite instability in colorectal cancer: a diagnostic study. Lancet Oncol. *22*, 132–141.

5. Coudray, N., Ocampo, P.S., Sakellaropoulos, T., Narula, N., Snuderl, M., Fenyö, D., Moreira, A.L., Razavian, N., and Tsirigos, A. (2018). Classification and mutation prediction from non-small cell lung cancer histopathology images using deep learning. Nat. Med. *24*, 1559–1567.

6. Holzinger, A., Biemann, C., Pattichis, C.S., and Kell, D.B. (2017). What Do We Need to Build Explainable AI Systems for the Medical Domain?. Preprint at arXiv. https://doi.org/10.48550/arXiv.1712.09923.

7. Lipton, Z.C. (2016). The Mythos of Model Interpretability. Preprint at arXiv. https://doi.org/10.48550/arXiv.1606.03490.

8. Kather, J.N., Pearson, A.T., Halama, N., Jäger, D., Krause, J., Loosen, S.H., Marx, A., Boor, P., Tacke, F., Neumann, U.P., et al. (2019). Deep learning can predict microsatellite instability directly from histology in gastrointestinal cancer. Nat. Med. *25*, 1054–1056.

9. Tschuchnig, M.E., Oostingh, G.J., and Gadermayr, M. (2020). Generative Adversarial Networks in Digital Pathology: A Survey on Trends and Future Potential. Patterns (New York, N.Y.) *1*, 100089.

10. Quiros, A.C., Murray-Smith, R., and Yuan, K. (2019). PathologyGAN: Learning Deep Representations of Cancer Tissue. Preprint at arXiv. https://doi.org/10.48550/arXiv.1907.02644.

11. Wei, T., Chen, D., Zhou, W., Liao, J., Zhang, W., Yuan, L., Hua, G., and Yu, N. (2022). E2Style: Improve the Efficiency and Effectiveness of StyleGAN Inversion. IEEE Trans. Image Process. *31*, 3267–3280. https://doi.org/10.1109/tip.2022.3167305.

12. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. (2016). Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. Preprint at arXiv. https://doi.org/10.48550/arXiv.1610.02391.

13. He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep Residual Learning for Image Recognition, pp. 770–778.

14. Zheng, X., Wang, R., Zhang, X., Sun, Y., Zhang, H., Zhao, Z., Zheng, Y., Luo, J., Zhang, J., Wu, H., et al. (2022). A deep learning model and human-machine fusion for prediction of EBV-associated gastric cancer from histopathology. Nat. Commun. *13*, 2790.

15. Woerl, A.C., Eckstein, M., Geiger, J., Wagner, D.C., Daher, T., Stenzel, P., Fernandez, A.,

Hartmann, A., Wand, M., Roth, W., and Foersch, S. (2020). Deep Learning Predicts Molecular Subtype of Muscle-invasive Bladder Cancer from Conventional Histopathological Slides. Eur. Urol. *78*, 256–264.

16. Iqbal, T., and Ali, H. (2018). Generative Adversarial Network for Medical Images (MI-GAN). J. Med. Syst. *42*, 231.

17. Krause, J., Grabsch, H.I., Kloor, M., Jendrusch, M., Echle, A., Buelow, R.D., Boor, P., Luedde, T., Brinker, T.J., Trautwein, C., et al. (2021). Deep learning detects genetic alterations in cancer histology generated by adversarial networks. J. Pathol. *254*, 70–79.

18. Cancer Genome Atlas Research Network (2014). Comprehensive molecular characterization of gastric adenocarcinoma. Nature *513*, 202–209. https://doi.org/10.1038/nature13480.

19. Lu, M.Y., Williamson, D.F.K., Chen, T.Y., Chen, R.J., Barbieri, M., and Mahmood, F. (2021). Data-efficient and weakly supervised computational pathology on whole-slide images. Nat. Biomed. Eng. *5*, 555–570.

https://doi.org/10.1038/s41551-020-00682-w.

20. Macenko, M., Niethammer, M., Marron, J.S., Borland, D., Woosley, J.T., Xiaojun, G., Schmitt, C., and Thomas, N.E. (2009). A Method for Normalizing Histology Slides for Quantitative Analysis, pp. 1107–1110.

21. Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., and Aila, T. (2020). Analyzing and Improving the Image Quality of StyleGAN, pp. 8107–8116.

22. Härkönen, E., Hertzmann, A., Lehtinen, J., and Paris, S. (2020). GANSpace: Discovering Interpretable GAN Controls. Preprint at arXiv. https://doi.org/10.48550/arXiv.2004.02546.

23. Abdal, R., Zhu, P., Mitra, N., and Wonka, P. (2020). StyleFlow: Attribute-Conditioned Exploration of StyleGAN-Generated Images Using Conditional Continuous Normalizing Flows. Preprint at arXiv. https://doi.org/10.48550/arXiv.2008.02401.

24. Alaluf, Y., Patashnik, O., and Cohen-Or, D. (2021). Only a Matter of Style: Age Transformation Using a Style-Based

Regression Model. Preprint at arXiv. https://doi.org/10.48550/arXiv.2102.02754.

25. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., and Wang, O. (2018). The Unreasonable Effectiveness of Deep Features as a Perceptual Metric, pp. 586–595.

26. Wright, L., and Demeure, N. (2021). Ranger21: A Synergistic Deep Learning Optimizer. Preprint at arXiv. https://doi.org/10.48550/arXiv.2106.13731.

27. Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., and Belongie, S. (2017). Feature Pyramid Networks for Object Detection, pp. 936–944.

28. Chong, M.J., Lee, H.-Y., and Forsyth, D. (2021). StyleGAN of All Trades: Image Manipulation with Only Pretrained StyleGAN. Preprint at arXiv. https://doi.org/10.48550/arXiv.2111.01619.

29. DeLong, E.R., DeLong, D.M., and Clarke-Pearson, D.L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics *44*, 837–845.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
Whole slide images	This paper	N/A
Source code	This paper	Data S1
Software and algorithms		
Python (version 3.8.13)	Python software	https://www.python.org/
Medcalc (vesion 15.2.2)	Medcalc software	https://www.medcalc.org/
IBM SPSS Statistics (version 20.0)	IBM SPSS Statistics software	https://www.ibm.com/cn-zh/products/spss-statistics

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources should be directed to and will be fulfilled by the Lead Contact, Muyan Cai (caimy@sysucc.org.cn).

Materials availability

This study did not generate new unique reagents.

Data and code availability

- All data reported in this paper will be shared by the lead contact upon reasonable request. TCGA slides can be obtained from the Genomic Data Commons portal (https://portal.gdc.cancer.gov/).
- The underlying code for this study is available in the Data S1 in a standalone ZIP file.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

Human subjects

To develop and validate MMRNet, we used three distinct cohorts: the Internal-STAD cohort from a single medical center, the MultiCenter-STAD cohort from multiple medical centers, and the TCGA-STAD public cohort from The Cancer Genome Atlas. Internal-STAD was used as the training dataset, which included all available slides from the patients with dMMR/MSI-H GC and randomly-selected slides from the pool of all patients with pMMR/non MSI-H GC in a single medical center. The MultiCenter-STAD included GC patients with available MMR/MSI status from multiple medical centers. For the TCGA-STAD, slides were obtained from the Genomic Data Commons portal (https://portal.gdc.cancer.gov/). Patients in our centers were selected based on the following criteria: (1) patients who underwent primary gastrectomy between January 1, 2014 and December 31, 2020; (2) having known MMR/MSI status; (3) having availability to the clinical data and H&E-stained tumor slides. We excluded patients who received preoperative therapy (e.g., neoadjuvant radiotherapy or chemotherapy), those with incomplete clinical information, and those with unqualified WSIs (e.g., slides out of focus, or obvious tissue folds). The study was approved by Sun Yat-sen University Cancer Center Ethics Committee and all experiments were carried out in accordance with existing guidelines of ethics committee. The average age within Internal-STAD was 61 years old, and the male-to-female ratio stood at 2.06. In the MultiCenter-STAD dataset, the average age was 58 years old, with a male-to-female ratio of 2.36. In the TCGA-STAD dataset, the average age was 66 years old, and the male-to-female ratio was 1.97. The sample size of this study was estimated through previously-published paper.[14] All patients in the Internal-STAD were included for MMRNet model training through random five-fold cross-validation. All patients in the MultiCenter-STAD and TCGA-STAD were included to externally validate the MMRNet.

METHOD DETAILS

Slide scanning

To obtain WSIs in SVS format, one or two representative H&E-stained tumor slides that contained a substantial portion of tumor tissue were scanned at 40× magnification (0.25 μm/pixel) using an Aperio AT2 scanner (Leica Biosystems; Wetzlar, Germany) from both the Internal-STAD and the MultiCenter-STAD.

Determination of MMR/MSI status

Immunohistochemistry (IHC) was used to confirm the MMR status of patients in Internal-STAD and MultiCenter-STAD. Using antibodies tagged with MMR proteins, IHC allowed for the visualization and localization of MMR proteins within the tissue. If any of the four major MMR proteins (MLH1, PMS2, MSH2, and MSH6) were absent, the sample was classified as dMMR/MSI-H tumors. Otherwise, they were classified as pMMR/non MSI-H type. In TCGA-STAD cohort, the MSI status of patients was determined through genetic sequencing, as previously published.[18]

Tumor detector

In this study, we aimed to automate the analysis of the MMR/MSI status from GC WSIs. To achieve this, we developed a tumor detector using the Internal-STAD and used it to automatically identify tumor regions in the external cohorts. The tumor detector was constructed using CLAM[19] in a weakly supervised pattern with a backbone of ImageNet pretrained ResNet50. A total of 1407 slides from Internal-STAD were included for training the tumor detector, which were divided into three parts for training, validation and testing set with a proportion of 8:1:1, and 10-fold cross-validation was performed. These slides comprised 202 dMMR/MSI-H GC slides, 1060 pMMR/non MSI-H GC slides, and 145 normal gastric tissue slides. Tiles with size of 256-by-256 pixels and a magnification ×10 were cropped from WSIs and fed to the tumor detector. ResNet50 extracted 2048-dimensional features for each tile, and the attention head of the tumor detector aggregated the features of all tiles in the same slide to obtain a slide-level tumor probability. Cross-entropy loss was used as the loss function. and the Adam optimizer was used for the optimization of the model parameters, with an initial learning rate of 0.0001 for 200 epochs. After 10 individual tumor detectors were developed, we used the 10 trained models for model ensemble prediction, which took the average of the scores of the 10 models as the final tumor score. For any tile from the external cohorts, 10 probability outputs of the classifiers were averaged as the final tumor score. Any region with the probability greater than 0.5 was taken as one tumor region of the slide.

MMRNet development and validation

To develop MMRNet, we extracted tiles from each slide according to their tumor score, ensuring that at least 40 tiles were extracted from each slide. To balance the data, we ultimately extracted a total of 219,143 dMMR/MSI-H tiles and 219,882 pMMR/non MSI-H tiles for model training, and 43,073 dMMR/MSI-H tiles and 43,371 pMMR/non MSI-H tiles for internal validation on Internal-STAD.

Similarly, we extracted 179,958 tiles from the MultiCenter-STAD cohort and 794,766 tiles from TCGA-STAD cohort. Each tile in the external cohorts was normalized using Macenko algorithm.[20] For MMRNet, we used ResNet18 with ImageNet pretrained weights as the backbone to predict the MSI/MMR status of each tumor tile. Data augmentation was applied to the training data, including random scaling and cropping, random horizontal flipping, random vertical flipping, random grayscale transformation, and random color transformations which include enhancements in brightness, contrast, hue, and saturation. Model finetuning adopted Adaptive Moment Estimation Weight (AdamW) as the optimizer with a batch size of 512 and an initial learning rate of 0.00001 for 300 epochs. To ensure model robustness, we performed a five-fold cross-validation for model training, which resulted in five well-trained sub-models. The MMRNet ensemble then averaged the outputs of these five sub-models to generate the final MMR/MSI status prediction for each slide. Additionally, we averaged the tile scores in the tumor area of a slide to obtain the MSI/MMR score for that slide.

Model interpretability

The StyleGAN[21] has been proven to generate highly realistic images with rich tissue structures. Due to the disentanglement of its latent space, several methods for latent space editing have been developed.[22,23] We aim to identify the morphological features that affect the classification outcomes of the MMRNet by manipulating the transformation from dMMR/MSI-H images to pMMR/non MSI-H images in the latent space. To achieve this, we implemented Alaluf et al.'s[24] approach and developed a conditional regression model, named MMRMapping. This model output a manipulation direction vector of the input image based on the category control signal, while preserving the overall structural similarity with an auxiliary similarity constraint.

We first trained a StyleGAN model on GC tiles in the Internal-STAD to obtain the latent space. We then performed 80,000 sampling iterations for each of the two categories, dMMR/MSI-H and pMMR/non MSI-H, to obtain their corresponding latent codes and generated images. The class labels of the images were reversed to obtain the control signals. The images and control signals were fed into the mapping network MMRMapping, which output the offset vectors. We added the offset vectors to the source latent codes to generate new pathological images. To guide the new generated images to have class labels opposite to the original labels, we utilized the MMRNet model as a supervised loss. In addition, we used the LILPS[25] loss to encourage the new images to have similar structures to the original images (Figure S9). To improve the training stability, we employed a cycle consistency pass to recover the source latent vector. The training process adopted the Ranger optimizer[26] with a learning rate of 1e-3 and a batch size of 8 images, and was performed for 30 epochs.

The MMRMapping was designed to manipulate the mapping direction of the input image based on a category signal. It consisted of a ResNet18 backbone[13] and a Feature Pyramid Network[27] module, which extracted multi-scale image features from coarse to fine. The input to the model included three RGB channels of the image and one channel for the category control signal, resulting in 512-dimensional features that were used to generate an offset vector. This vector, when added to the latent code corresponding to the input image itself, allowed for regeneration of a new image with a class opposite to that of the input image, providing a means to summarize the morphological features of dMMR/MSI-H GC (Figure S10).

To focus on local areas closely related to classification, the CAM-blending algorithm was developed for the inference stage. Although a similarity constraint was applied during the training process of MMRMapping, the global characteristic of StyleGAN's latent code makes it difficult to manipulate local areas. The CAM-blending combined CAM and alpha-blending[28] to generate images that highlighted the high attention areas of MMRNet. The original latent codes and the offset vector generated by MMRMapping were used as inputs of StyleGAN, producing a feature output in the network. Grad-CAM9[12] was then used to generate CAM during MMRNet classification, which was employed to linearly combine the output features of StyleGAN, resulting in a synthesized image that seamlessly blended the significant morphological features of dMMR/MSI-H GC while inhibiting transformation in low-attention areas (Figure S11).

Identification and validation of the morphological features related to dMMR/MSI-H GC

Two expert pathologists reviewed 10,000 patch groups generated by StyleGAN to identify the recognizable morphological changes. Each patch group contained five H&E-stained patches and their corresponding heatmaps. MMRNet predictive scores of the patches gradually decreased as MMRMapping manipulated the offset vector of high-attention regions. Pathologists recorded only the morphological features that gradually disappeared in the high-attention regions of the heatmap. To validate these features, two senior pathologists reviewed the slides from two external cohorts and recorded whether the slide possessed the summarized morphological features. Morphological features of dMMR/MSI-H slides were then compared to those of pMMR/non MSI-H slides.

Reader study

To evaluate the impact of summarized morphological features on pathologist performance, a reader study was conducted with the participation of four junior pathologists. Before final evaluation, a visual presentation illustrating distinct morphological features characteristic of dMMR/MSI-H GC as identified by expert pathologists was served as valuable educational tools, aiding junior pathologists in recognizing and comprehending these intricate tumor subtypes. Then junior pathologists reviewed a testing set of 60 WSIs randomly selected from MultiCenter-STAD and TCGA-STAD datasets. The testing set comprised 30 WSIs with dMMR/MSI-H GC and 30 WSIs with pMMR/non MSI-H GC, representing a total of 60 patients. All pathologists were blinded to all clinical information, including the dMMR/MSI-H and pMMR/non MSI-H ratio in the dataset. For each WSI, they assessed whether they thought the cancer could be classified as dMMR/MSI-H or pMMR/non MSI-H, based on their review of the WSI. They were then given the option to change their initial assessment after learning the summarized morphological features.

QUANTIFICATION AND STATISTICAL ANALYSIS

The study conducted various statistical analyses to evaluate the performance of the MMRNet model and the impact of the summarized morphological features on pathologist diagnosis. The AUROCs of MMRNet were calculated by Delong et al.[29] using the ground-truth MMR/MSI status as the reference standard. This threshold was predefined in the Internal-STAD and prior to the assessment on external datasets. The morphological features of dMMR/MSI-H cases were compared with those of pMMR/non MSI-H cases using the Chi-square test. To compare the diagnostic performances of pathologists before and after they learned the summarized morphological features, the differences of their performance were compared with Delong et al. method. A difference was considered significant when the *P* value from a two-tailed test was less than 0.05. IBM SPSS Statistics (version 20.0) and Medcalc (vesion 15.2.2) were used for statistical analysis, while Python (version 3.8.13) and the deep learning platform PyTroch (version 1.11.0) were performed to build the model and analyze the data.