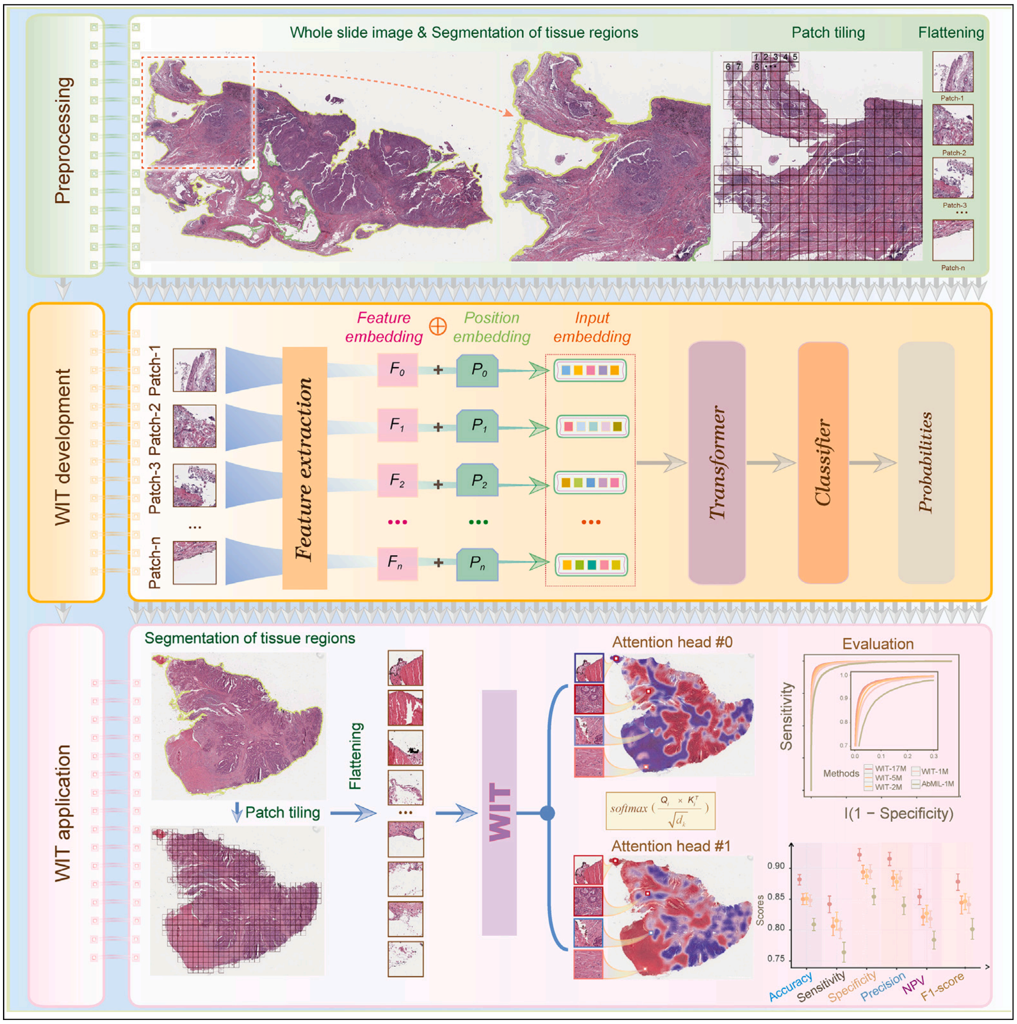
iScience

CellPress
OPEN ACCESS

Article

An efficient context-aware approach for whole-slide image classification



Hongru Shen, Jianghua Wu, Xilin Shen, ..., Yan Sun, Kexin Chen, Xiangchun Li

chenkexin@tmu.edu.cn (K.C.)
lixiangchun2014@foxmail.com (X.L.)

Highlights

WIT aggregates representations of all image patches for slide-level classification

WIT achieves high accuracy in the detection of 32 cancer types and diagnosis of cancer

Saliency maps obtained from WIT are visually interpretable

Shen et al., iScience *26*, 108175
December 15, 2023 © 2023 The Authors.
https://doi.org/10.1016/j.isci.2023.108175

iScience

Article

An efficient context-aware approach for whole-slide image classification

Hongru Shen,[1,6] Jianghua Wu,[2,6] Xilin Shen,[1] Jiani Hu,[1] Jilei Liu,[1] Qiang Zhang,[3] Yan Sun,[4] Kexin Chen,[5,*] and Xiangchun Li[1,7,*]

SUMMARY

Computational pathology for gigapixel whole-slide images (WSIs) at slide level is helpful in disease diagnosis and remains challenging. We propose a context-aware approach termed WSI inspection via transformer (WIT) for slide-level classification via holistically modeling dependencies among patches on WSI. WIT automatically learns feature representation of WSI by aggregating features of all image patches. We evaluate classification performance of WIT and state-of-the-art baseline method. WIT achieved an accuracy of 82.1% (95% CI, 80.7%–83.3%) in the detection of 32 cancer types on the TCGA dataset, 0.918 (0.910–0.925) in diagnosis of cancer on the CPTAC dataset, and 0.882 (0.87–0.890) in the diagnosis of prostate cancer from needle biopsy slide, outperforming the baseline by 31.6%, 5.4%, and 9.3%, respectively. WIT can pinpoint the WSI regions that are most influential for its decision. WIT represents a new paradigm for computational pathology, facilitating the development of digital pathology tools.

INTRODUCTION

The development of digital pathology leads to accumulation of large-scale whole-slide imaging data, laying the foundation of big data for computational pathology. Rich morphological features buried in whole-slide image (WSI) provide diagnostic information of the disease and offer guidance on the decision for treatment. Advances in deep learning algorithms enable the analyses of gigapixel WSIs at scale for disease diagnosis,[1–3] prognosis,[4–7] and treatment selection.[8,9]

Deep learning approaches have achieved human-level performance in recognizing natural images in the ImageNet competition.[10–13] However, automatic recognition of WSI remains challenging due to the super-high spatial resolution of WSI as compared with images from ImageNet.[10] To address this challenge, researchers divided WSI into small image patches and subsequently aggregated the features of image patches to obtain slide-level features.[5,14–17] For example, Campanella and colleagues used standard multiple-instance learning (MIL) to diagnose prostate cancer, basal cell carcinoma, and auxiliary lymph node metastasis of breast cancer by first ranking image patches with regard to slide-level labels and using the most relevant image patch for slide-level classification.[1] Lu and colleagues developed a data-efficient weakly supervised approach[18] for slide-level classification using attention-based pooling[19] of all image patches instead of the most relevant patch used by standard MIL.[1] Based on this approach, Lu and colleagues introduced tumor origin assessment via deep learning (TOAD) to predict tissue-of-origins for cancer of unknown primary.[20] Meanwhile, this attention-based MIL method has been utilized for addressing the diagnostic tasks for cardiac allograft rejection screening in WSIs[8] and prognostic prediction by fusing WSIs with different modalities of genomic data.[4] Apart from these diagnostic endeavors, analyses of large-scale WSIs have been proved to be feasible for the prediction of genetic markers. Coudray and colleagues reported a deep-learning-based approach for predicting somatic mutations in canonical driver genes for lung cancer via averaging the probabilities of image patches or counting the percentage of image patches classified as positive.[15] In addition, multiple studies reported that micro-satellite instability can be predicted from WSIs in gastrointestinal cancer,[21] colorectal cancer,[22–24] and endometrial carcinoma.[25]

The transformer architecture designed for natural language understanding can capture long-range dependencies among different entities.[26] Transformer-based language architectures have achieved superior performance in various language understanding tasks.[26–28] The self-attention operation is the key module underlying the success of transformer in that it captures dependencies in the input.[26] Although

[1] Tianjin Cancer Institute, Tianjin's Clinical Research Center for Cancer, National Clinical Research Center for Cancer, Tianjin Medical University Cancer Institute and Hospital, Tianjin Medical University, Tianjin, China
[2] Department of Pathology, Peking University Cancer Hospital & Institute, Key Laboratory of Carcinogenesis and Translational Research (Ministry of Education), Beijing, China
[3] Department of Maxillofacial and Otorhinolaryngology Oncology, Tianjin's Clinical Research Center for Cancer, National Clinical Research Center for Cancer, Tianjin Medical University Cancer Institute and Hospital, Tianjin Medical University, Tianjin, China
[4] Department of Pathology, Tianjin's Clinical Research Center for Cancer, Key Laboratory of Cancer Immunology and Biotherapy, National Clinical Research Center for Cancer, Tianjin Cancer Institute and Hospital, Tianjin Medical University, Tianjin, China
[5] Department of Epidemiology and Biostatistics, Tianjin's Clinical Research Center for Cancer, Key Laboratory of Molecular Cancer Epidemiology of Tianjin, National Clinical Research Center for Cancer, Tianjin Medical University Cancer Institute and Hospital, Tianjin Medical University, Tianjin, China
[6] These authors contributed equally
[7] Lead contact
*Correspondence: chenkexin@tmu.edu.cn (K.C.), lixiangchun2014@foxmail.com (X.L.)

https://doi.org/10.1016/j.isci.2023.108175

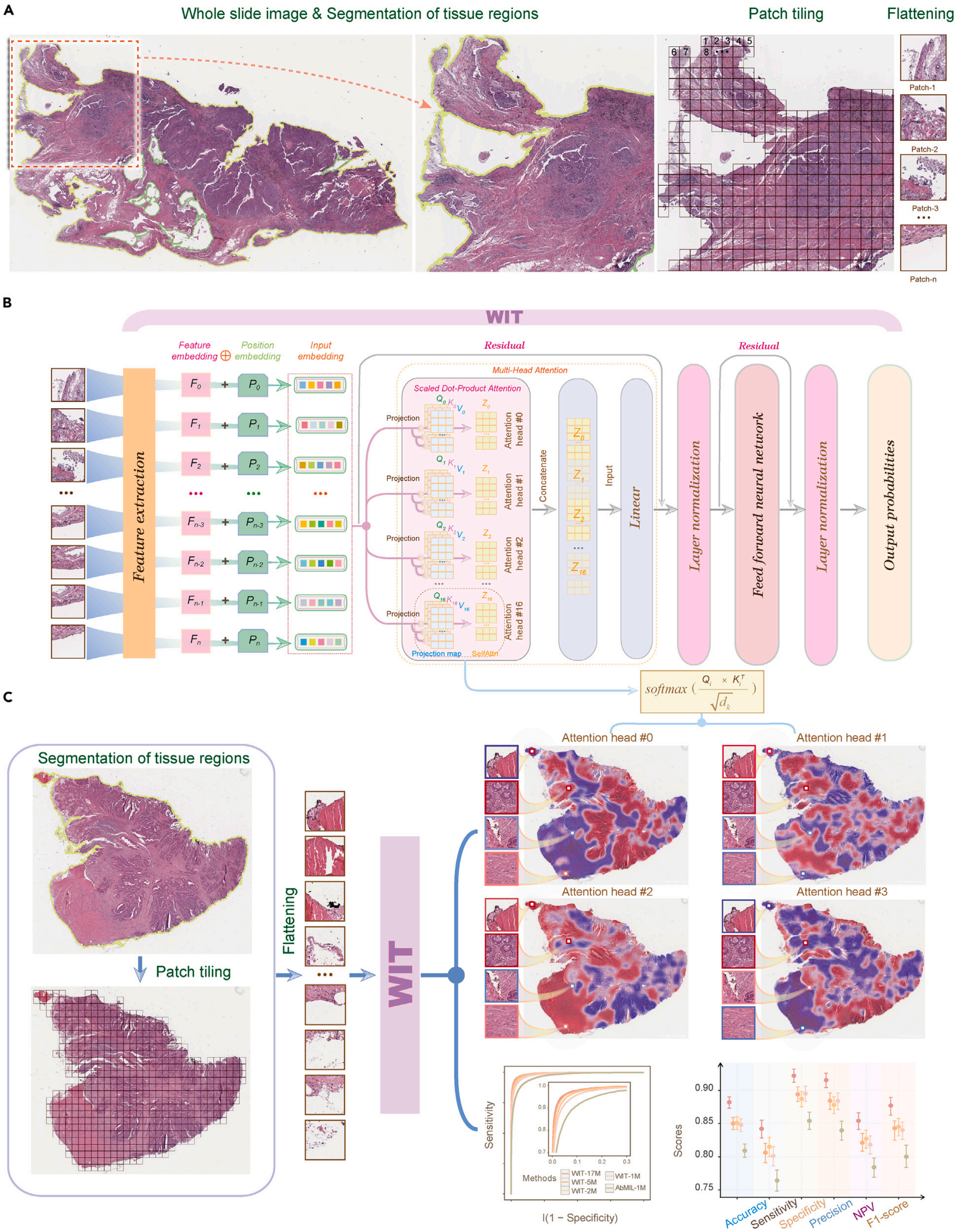

A
Whole slide image & Segmentation of tissue regions
Patch tiling
Flattening
1 2 3 4 5
6 7 8
Patch-1
Patch-2
Patch-3
Patch-n
B
WIT
Feature embedding ⊕ Position embedding
Input embedding
Feature extraction
F_0 + P_0
F_1 + P_1
F_2 + P_2
F_{n-3} + P_{n-3}
F_{n-2} + P_{n-2}
F_{n-1} + P_{n-1}
F_n + P_n
Residual
Residual
Multi-Head Attention
Scaled Dot-Product Attention
Projection
Q_0 K_0 V_0
Z_0
Attention head #0
Q_1 K_1 V_1
Z_1
Attention head #1
Q_2 K_2 V_2
Z_2
Attention head #2
Q_{16} K_{16} V_{16}
Z_{16}
Attention head #16
Projection map
SelfAttn
Concatenate
Input
Linear
Layer normalization
Feed forward neural network
Layer normalization
Output probabilities
$softmax\left(\frac{Q_i \times K_i^T}{\sqrt{d_k}}\right)$
C
Segmentation of tissue regions
Patch tiling
Flattening
WIT
Attention head #0
Attention head #1
Attention head #2
Attention head #3
Sensitivity
l(1 − Specificity)
1.0
0.9
0.8
0.7
0.0
0.1
0.2
0.3
Methods
WIT-17M
WIT-5M
WIT-2M
WIT-1M
AbMIL-1M
Scores
0.90
0.85
0.80
0.75
Accuracy
Sensitivity
Specificity
Precision
NPV
F1-score

Figure 1. A flowchart illustrating the framework of WIT

(A) Illustration of the preprocessing steps: segmentation of tissue regions, patch tiling and flattening.

(B) The architecture of WIT.

(C) Evaluation of WIT for classification and model interpretability. WSI, Whole Slide Image; AbMIL, Attention-based Multiple Instance Learning.

it was proposed for language understanding, transformer is inherently task-agnostic. It has been widely adopted or revised for image recognition. Vision transformer (ViT) is a direct adoption of transformer for image classification by splitting image into multiple patches and taking the flatten image patches as input.[26] Thereafter, ViT-based architectures have been widely used in medical imaging analyses.[29–31]

Inspired by the success of transformer-based natural language understanding[32,33] and image recognition,[27] we present an approach called WSI inspection via transformer (WIT) for slide-level classification via holistically modeling dependencies among patches on the WSI. WIT takes as input the features of image patches that were extracted with an image model pretrained on ImageNet.[34] We collected a total number of 22,457 WSIs from TCGA, CPTAC, and PANDA projects to develop and systematically evaluate WIT for detection of 32 cancer types and diagnosis of cancer. The TCGA consists of 11,623 WSIs covering 32 cancer types. The CPTAC dataset includes 3,414 WSIs from cancer patients and 1,638 WSIs from non-cancer controls. The PANDA dataset consists of 5,782 needle biopsy slides; 2,891 of them are prostate cancers and rest are non-cancer controls. WIT achieved an accuracy of 82.1% in the detection of 32 cancer types on the TCGA dataset, 91.8% in diagnosis of cancer on the CPTAC dataset, and 88.2% on the PANDA dataset, outperforming the attention-based MIL baseline by 31.6%, 5.4%, and 9.3%, respectively. WIT can pinpoint the WSI regions that are most influential for its decision. WIT represents a new paradigm for computational pathology. It will facilitate the development of assistive tools for digital pathology.

RESULTS

An overview of WIT

The procedures to develop WIT includes WSI segmentation and tiling, model development, and evaluation (Figure 1). Firstly, we segmented the WSI to identify tissue regions and subsequently tiled WSI into patches of 256 × 256 pixels (Figure 1A). WIT takes these flattened image patches as input. We used a pretrained model to extract a feature with 1,024 dimensions for each image patch (See STAR methods). Meanwhile, the position embeddings of image patches on that WSI along with their extracted feature were fed into a transformer block. The transformer block consists of a multi-headed self-attention module and point-wise feed-forward neural network. Residual connection is employed around these two sub-modules, followed by layer normalization[26] (Figure 1B). The multi-headed self-attention module learns the dependencies among different image patches and the influence of each patch on the output, such as slide labels (Figure 1B). WIT was evaluated for its capacity in slide classification and localization of image patches that exhibit significant association with slide labels (Figure 1C).

High performance of WIT in tissue-of-origin localization

We systematically evaluated the classification performance of WIT on The Cancer Genome Atlas (TCGA) dataset for tissue-of-origin localization via 5-fold cross-validation (See STAR methods). The TCGA dataset consists of 11,623 formalin-fixed paraffin-embedded WSIs from 9,565 individuals covering 32 cancer types (Table S1). We examined classification performance of WIT with varying parameters such as 1, 2, 5, and 17 megabytes (Table S2). We used the attention-based MIL model as the baseline model for comparison. The baseline possesses model parameters of 1 megabyte.

The accuracy of WIT was increasing with model size. Its top-1 accuracy ranged from 73.1% (95% confidence interval [CI], 72.5%–73.9%) for WIT-1Mb to 82.1% (80.7%–83.3%) for WIT-17Mb, whereas the baseline had a top-1 accuracy of 64.2% (60.6%–66.0%) (Figure 2A). Top-2 and top-3 accuracies exhibited the same trend as top-1 accuracy (Figure 2A; Table S3). Meanwhile, the micro-average AUROC of four WIT models were also higher than the baseline model (Figure 2B). WIT-17Mb achieved high performance in localization of 32 cancer types with respect to precision and recall rate (Figure 2C). WIT-17M achieved an average precision of 77.3% and recall rate of 75.6%, outperforming the baseline method by 29.5% and 37.5%, respectively. The confusion matrix of the baseline method was shown in Figure S1D. WIT of different model size also had higher performance as compared with the baseline method when stratified by cancer types (Figure 2D; Tables S4–S6). In addition, the F1 scores achieved by different WIT models are higher than the baseline method (Figure 2E; Table S7). For example, WIT-1M had an average F1 score of 0.618 versus 0.554 as obtained by the baseline method, albeit WIT-1M and the baseline method had comparable model size.

High performance of WIT in cancer diagnosis

WIT achieved high classification performance in the diagnosis of cancer on the CPTAC and PANDA datasets (See STAR methods). The CPTAC dataset consists of 5,052 formalin-fixed paraffin-embedded WSIs from 1,330 individuals (Table S8). The PANDA dataset consists of 5,782 prostate WSIs subjected to needle biopsies.[35]

On the CPTAC dataset, WIT models achieved AUROCs ranging from 0.941 (95% CI, 0.934–0.949) to 0.953 (0.946–0.960), whereas the baseline model achieved an AUROC of 0.931 (0.931–0.969) (Figure 3A). WIT-17Mb achieved an accuracy of 0.918 (0.910–0.925) as compared with WIT models of smaller sizes as well as the baseline model. Similar trends were observed with respect to other classification metrics (Figures 3B and 3C; Table S9). On the PANDA dataset, WIT-17Mb achieved the significantly higher AUROC as compared with WIT of smaller sizes and the baseline model (DeLong's test, all adjusted p values <2.2e-16, Figure 3D). Classification metrics such as accuracy, sensitivity, specificity, precision, negative predictive value, and F1 score achieved by WIT-17Mb were also significantly higher than the other models (Figures 3E and 3F; Table S10).

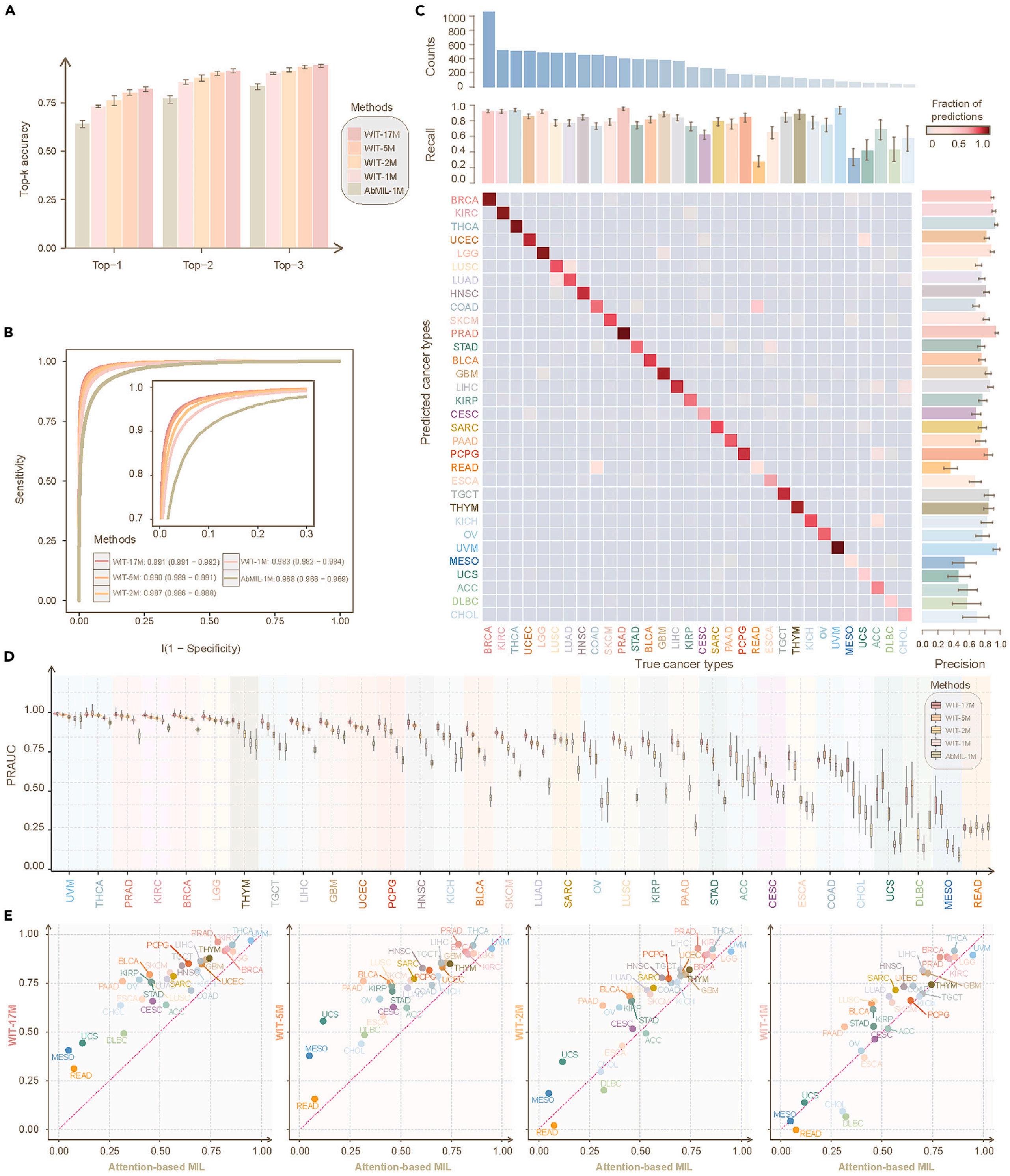


Figure 2. The classification performance of WIT in localization of tissue origins for 32 cancer types on TCGA dataset

(A) Top-*K* accuracy for localization of tumor origins, $K \in \{1, 2, 3\}$.

(B) Micro-average area under the receiver operating curve.

(C) Patient-level performance from 5-fold cross-validation. Per origin count, precision, and recall rate are plotted next to the confusion matrix. The columns represent the true origin of the tumor, and rows represent the prediction by the WIT model.

(D) Area under the precision-recall curve (PRAUC) stratified by cancer types.

(E) Scatterplots of F1 scores between different models. AbMIL, attention-based multiple instance learning.

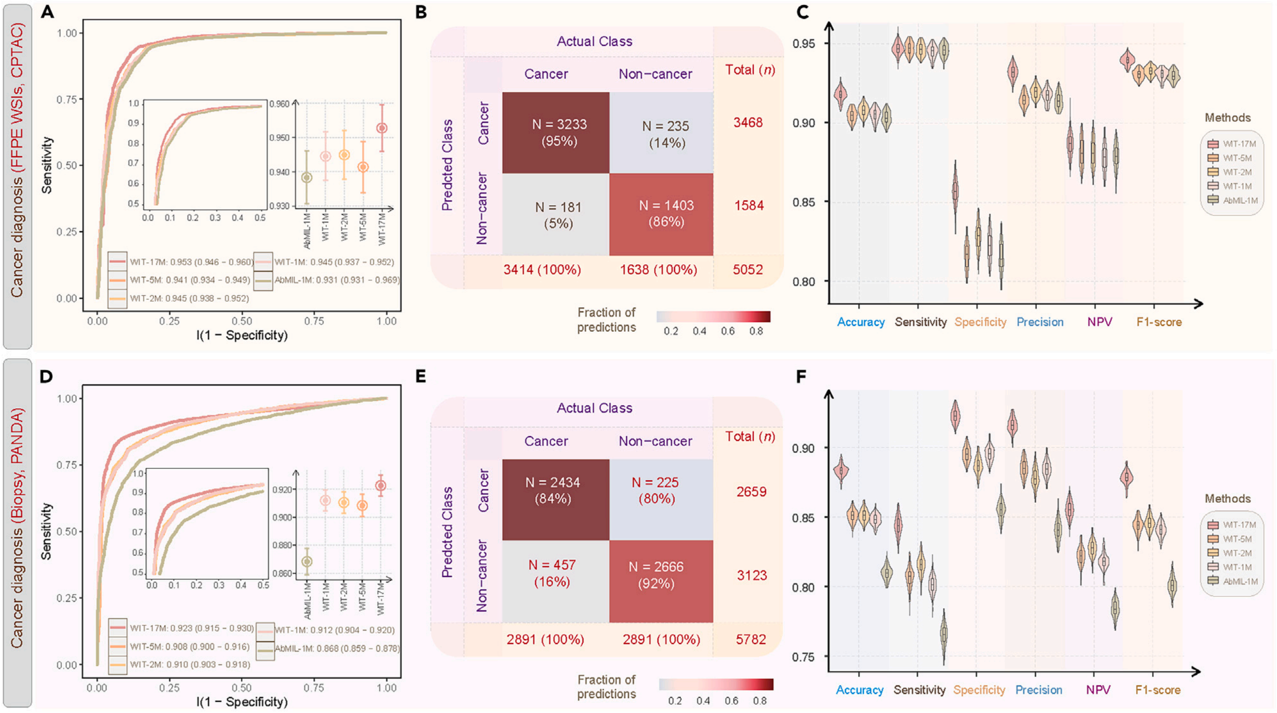


Figure 3. The classification performance of WIT in the diagnosis of cancer on CPTAC and PANDA datasets

(A and D) The receiver operating curves and area under the curves.

(B and E) Confusion matrices.

(C and F) Classification metrics of accuracy, sensitivity, specificity, precision, negative predictive value (NPV), and F1-score. AbMIL, attention-based multiple instance learning.

Model interpretability

The multi-headed self-attention modules in WIT measure the association between the classification representation and each image patch. Therefore, the attention scores can be interpreted as the association between each image patch and the classification output. We converted attention scores derived from WIT into human-interpretable heatmaps, which highlights importance of WSI regions for prediction (See STAR methods). In localization of 32 cancer types, WIT captures tumor regions that are considered to be morphology of different cancer types by pathologists in lung adenocarcinoma (LUAD, Figure 4A), rectum adenocarcinoma (Figure 4B), pancreatic adenocarcinoma (Figure 4C), and uterine corpus endometrial carcinoma (Figure 4D). For example, WIT identifies micropapillary tufts forming florets structure as strong evidence in detection of LUAD (Figure 4A). In the diagnosis of cancer, WIT pinpoints the tumor regions of non-keratinizing squamous cells with solid pattern in lung squamous cell carcinoma (Figure 4E) and confluent glandular and cribriform structure in UCEC (Figure 4F). In addition, WIT is able to identify prostate adenocarcinoma (Figure 4G and 4H) and a cluster of small poorly formed glands (Figure 4G) from needle biopsy. We provided visualization of attention maps for a number of slides for exploration purpose in our interactive website (https://deeplearningplus.github.io/WIT-attention-maps/).

DISCUSSION

In our study, we proposed a context-aware deep learning approach WIT for slide-level localization of tumor origins and diagnosis of cancer from WSIs. WIT outperformed the attention-based MIL[20] baseline by significant marginals across all classification tasks evaluated, especially in the detection of 32 cancer types where WIT achieved a micro-average area under the receiver operating curve (AUROC) of 0.991 (0.991–0.992) versus 0.968 (0.966–0.969) as obtained by the baseline method.

The high performance of WIT can be attributed to its context-aware ability to learn the potential nonlinear associations among image patches, whereas the baseline method treats different image patches as independent instances. As WIT was built upon transformer,[36] the multi-headed self-attention module in transformer enables WIT to learn interrelation of patches in different subspaces, whereas attention-based multiple-instance learning (MIL) is designed to aggregate multiple instances independently. Attention-based MIL methods have been widely and successfully adopted in addressing the challenges of computational pathology such as CLAM,[18] TOAD,[20] and CRANE.[8] WIT has the advantage of CLAM and TOAD in that it uses only the slide-level labels without any manual annotation. However, both CLAM and TOAD share the common limitations of MIL-based approaches[37] in that they are context-independent but not context-aware.

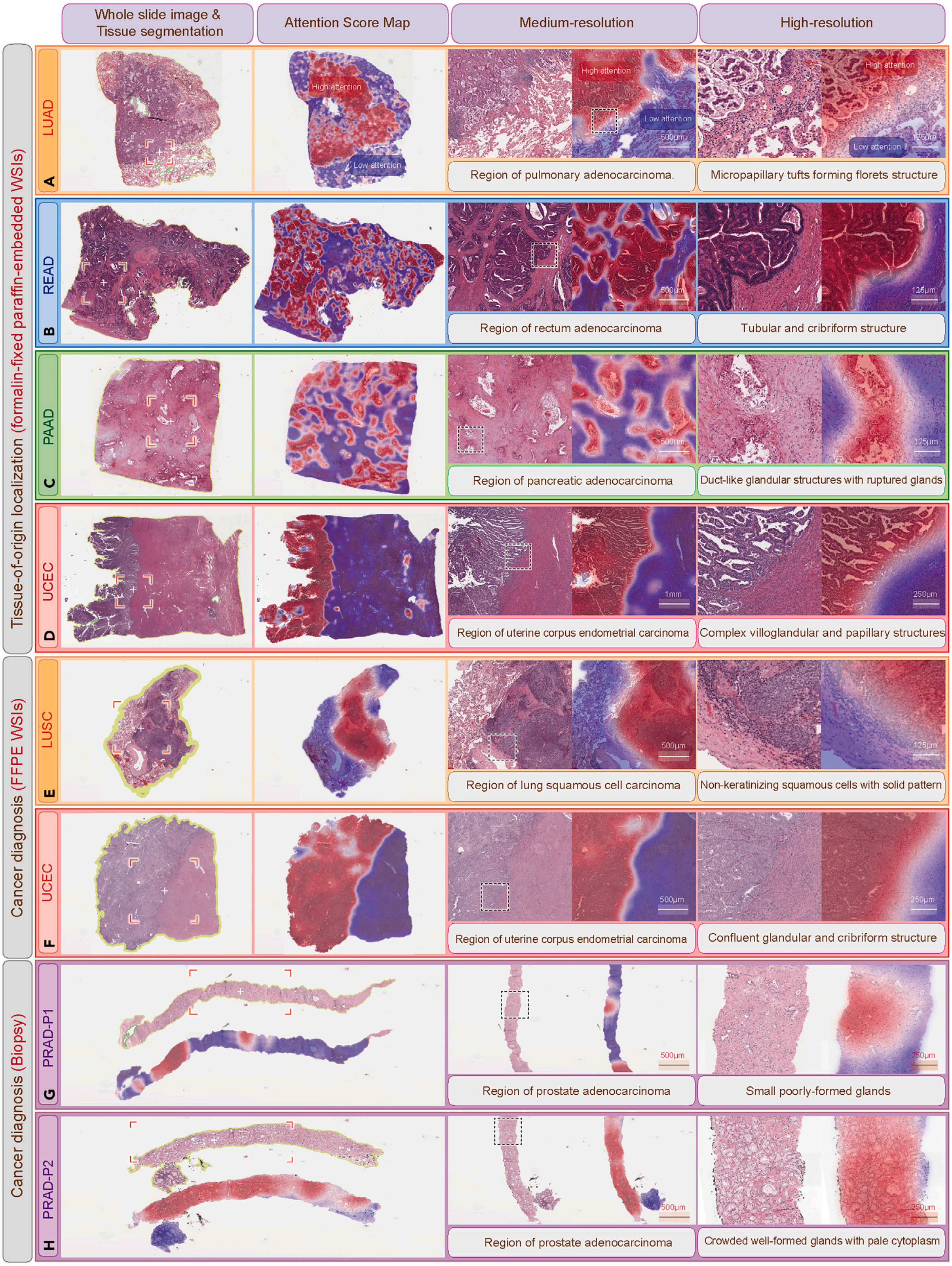

Whole slide image & Tissue segmentation
Attention Score Map
Medium-resolution
High-resolution
Tissue-of-origin localization (formalin-fixed paraffin-embedded WSIs)
Cancer diagnosis (FFPE WSIs)
Cancer diagnosis (Biopsy)
A LUAD
High attention
Low attention
500μm
125μm
Region of pulmonary adenocarcinoma.
Micropapillary tufts forming florets structure
B READ
500μm
125μm
Region of rectum adenocarcinoma
Tubular and cribriform structure
C PAAD
500μm
125μm
Region of pancreatic adenocarcinoma
Duct-like glandular structures with ruptured glands
D UCEC
1mm
250μm
Region of uterine corpus endometrial carcinoma
Complex villoglandular and papillary structures
E LUSC
500μm
125μm
Region of lung squamous cell carcinoma
Non-keratinizing squamous cells with solid pattern
F UCEC
500μm
250μm
Region of uterine corpus endometrial carcinoma
Confluent glandular and cribriform structure
G PRAD-P1
500μm
250μm
Region of prostate adenocarcinoma
Small poorly-formed glands
H PRAD-P2
500μm
250μm
Region of prostate adenocarcinoma
Crowded well-formed glands with pale cytoplasm

Figure 4. Attention maps of WIT for interpretability in localization of tissue origins and diagnosis of cancer from FFPE WSIs and biopsy

Boxes highlight the typical morphologic features corresponding to the textual description. The interactive visualization is available at https://deeplearningplus.github.io/WIT-attention-maps/.

As compared with TOAD developed in the previous study, our method has fine-grainer classification. Our method performs classification for 32 cancer types, whereas TOAD performs classification for 18 cancer types. TOAD did not include MESO and DLBCL and did not distinguish between READ and COAD; LUSC and LUAD; and KICH, KIRC, and KIRP. In contrast, our method treats each of these cancer subtypes as different classes.

Better performance for UVM, THCA, and PRAD when compared with DLBC, MESO, and READ is related to their morphological features. For example, UVM is characterized by well distinctive features such as ciliary body location, diffuse-type tumor, ring melanoma of the iris, presence of vascular mimickers, and extraocular extension.[38] THCA presents with a papillary pattern or a follicular pattern with or without thyroid colloid.[39] PRAD is characterized with perineural invasion, glomerulations, and mucinous fibroplasia (also known as collagenous micronodule).[40] These features of UVM, THCA, and PRAD are separately unique and predominantly different from other cancer types. Conversely, DLBC, MESO, and READ are more complex and challenging for accurate diagnosis. DLBCLs are characterized by partial or complete effacement of the normal architecture (nodal or extranodal) by medium- to large-sized lymphoid cells with vesicular chromatin. These features necessitate immunohistochemical staining in clinical setting for confirmatory diagnosis.[41] Mesothelioma cells were morphologically diverse. It is difficult to distinguish between epithelioid mesothelioma and metastatic carcinoma.[42] READs are characterized by glandular tubular or diffuse nests depending on its differentiation. These features of tumor cells or structure are not specific among these tumors, and it is difficult to distinguish these tumors from STAD and COAD.

WIT has several specific advantages. First, WIT can be easily scaled into models of different sizes. Large model has better classification performance as compared with smaller ones. However, the high performance of different WIT models cannot be merely attributed to their model sizes as compared with the attention-based MIL baseline. For example, in the detection of 32 cancer types, WIT-1Mb achieved significantly higher overall accuracy in comparison to the baseline method [73.1% (95% CI, 72.5%–73.9%) versus 64.2% (62.4%–66.0%)] although their model sizes are comparable. Therefore, the high performance of WIT is likely due to its ability to take into account nonlinear associations among all image patches. Besides, overall accuracy is steadily increasing with model size (Table S2). Second, WIT is data-efficient in that we extracted image patches at ×20 magnification instead of full magnification. In this scenario, the 16 terabytes of TCGA WSI dataset were converted into a dataset of 200 gigabytes, enabling fast experimentation. Third, multi-head attentions used by WIT enable model interpretability from different feature representation subspaces, allowing for different morphological features to be identified by different attention heads. For example, we observed that one attention head of WIT identified micropapillary tufts forming floret structure as strong evidence for lung adenocarcinoma (Figure 4A), whereas the other heads pay attention to different tissue structure such as normal pulmonary alveoli (Figure S2).

Conclusion

Weakly supervised learning such as MIL-based approaches have been successfully applied in addressing the challenges of computational pathology. However, their limitations are apparent in that they treat instances independently. Here, we addressed this challenge by presenting WIT—a deep learning method based on transformer for learning feature presentation of whole slide by taking into account nonlinear associations among image patches. WIT will facilitate adoption of deep-learning-based solution and enable knowledge discovery in computational pathology.

Limitations of the study

However, WIT was not without limitations. We used the ResNet50 model[34] pretrained on the ImageNet dataset as feature extractor for image patches of WSI. The ImageNet is a collection of natural scene images. Therefore, it is definitely suboptimal by using this pretrained ResNet50 model[34] in characterizing image patches clipped from WSIs. This strategy was also adopted by CLAM, TOAD, and CRANE. Pretraining the feature extractor on image patches of WSIs may have the potential to improve the performance of WIT and all MIL-based methods. However, this will drastically increase the computational resources. We will address this issue in our future study. In addition, the 2D spatial dependencies among image patches is lost, as WIT accepts flattened patches as input. Addressing this drawback with multi-dimensional transformers such as axial attention[43] will improve the performance of WIT.

STAR★METHODS

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

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2023.108175.

ACKNOWLEDGMENTS

We are grateful for researchers for their generosity to make their data publicly available. This work was supported by the National Natural Science Foundation of China (Grant No. 32270688 and 31801117 to X.L. and 82073287 to Q.Z.), National Key Research and Development Program of China (Grant No. 2021YFC2500400 to K.C.), and Program for Changjiang Scholars and Innovative Research Team in University in China (Grant No. IRT_14R40 to K.C.). This work was funded by Tianjin Key Medical Discipline (Specialty) Construction Project (TJYXZDXK-009A).

AUTHOR CONTRIBUTIONS

Xiangchun Li and Kexin Chen designed and supervised the study; Xiangchun Li and Hongru Shen performed data analysis and wrote the manuscript; Xiangchun Li developed the model; Jianghua Wu interpreted the whole-slide image data. Xiangchun Li, Hongru Shen, Xilin Shen, Jiani Hu, Jilei Liu, and Qiang Zhang collected data; Yan Sun provided comments on the results. Hongru Shen, Xiangchun Li, and Kexin Chen revised the manuscript.

DECLARATION OF INTERESTS

The authors declare that they have no conflict of interest.

Received: March 20, 2023
Revised: August 29, 2023
Accepted: October 8, 2023
Published: October 12, 2023

REFERENCES

1. Campanella, G., Hanna, M.G., Geneslaw, L., Miraflor, A., Werneck Krauss Silva, V., Busam, K.J., Brogi, E., Reuter, V.E., Klimstra, D.S., and Fuchs, T.J. (2019). Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. Nat. Med. *25*, 1301–1309. https://doi.org/10.1038/s41591-019-0508-1.
2. Ström, P., Kartasalo, K., Olsson, H., Solorzano, L., Delahunt, B., Berney, D.M., Bostwick, D.G., Evans, A.J., Grignon, D.J., Humphrey, P.A., et al. (2020). Artificial intelligence for diagnosis and grading of prostate cancer in biopsies: a population-based, diagnostic study. Lancet Oncol. *21*, 222–232. https://doi.org/10.1016/S1470-2045(19)30738-7.
3. Kotei, E., and Thirunavukarasu, R. (2022). Computational techniques for the automated detection of mycobacterium tuberculosis from digitized sputum smear microscopic images: A systematic review. Prog. Biophys. Mol. Biol. *171*, 4–16. https://doi.org/10.1016/j.pbiomolbio.2022.03.004.
4. Chen, R.J., Lu, M.Y., Williamson, D.F.K., Chen, T.Y., Lipkova, J., Noor, Z., Shaban, M., Shady, M., Williams, M., Joo, B., and Mahmood, F. (2022). Pan-cancer integrative histology-genomic analysis via multimodal deep learning. Cancer Cell *40*, 865–878.e6. https://doi.org/10.1016/j.ccell.2022.07.004.
5. Mobadersany, P., Yousefi, S., Amgad, M., Gutman, D.A., Barnholtz-Sloan, J.S., Velázquez Vega, J.E., Brat, D.J., and Cooper, L.A.D. (2018). Predicting cancer outcomes from histology and genomics using convolutional networks. Proc. Natl. Acad. Sci. USA *115*, E2970–E2979. https://doi.org/10.1073/pnas.1717139115.
6. Chen, R.J., Lu, M.Y., Wang, J., Williamson, D.F.K., Rodig, S.J., Lindeman, N.I., and Mahmood, F. (2022). Pathomic Fusion: An Integrated Framework for Fusing Histopathology and Genomic Features for Cancer Diagnosis and Prognosis. IEEE Trans. Med. Imaging *41*, 757–770. https://doi.org/10.1109/TMI.2020.3021387.
7. Courtiol, P., Maussion, C., Moarii, M., Pronier, E., Pilcer, S., Sefta, M., Manceron, P., Toldo, S., Zaslavskiy, M., Le Stang, N., et al. (2019). Deep learning-based classification of mesothelioma improves prediction of patient outcome. Nat. Med. *25*, 1519–1525. https://doi.org/10.1038/s41591-019-0583-3.
8. Lipkova, J., Chen, T.Y., Lu, M.Y., Chen, R.J., Shady, M., Williams, M., Wang, J., Noor, Z., Mitchell, R.N., Turan, M., et al. (2022). Deep learning-enabled assessment of cardiac allograft rejection from endomyocardial biopsies. Nat. Med. *28*, 575–582. https://doi.org/10.1038/s41591-022-01709-2.
9. Shamai, G., Livne, A., Polónia, A., Sabo, E., Cretu, A., Bar-Sela, G., and Kimmel, R. (2022). Deep learning-based image analysis predicts PD-L1 status from H&E-stained histopathology images in breast cancer. Nat. Commun. *13*, 6753. https://doi.org/10.1038/s41467-022-34275-9.
10. Krizhevsky, A., Sutskever, I., and Hinton, G.E. (2017). Imagenet classification with deep convolutional neural networks. Commun. ACM 60, 84–90.
11. Huang, G., Liu, Z., Van Der Maaten, L., and Weinberger, K.Q. (2017). Densely Connected Convolutional Networks, pp. 4700–4708.
12. He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep Residual Learning for Image Recognition, pp. 770–778.

13. Simonyan, K., and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. Preprint at arXiv. https://doi.org/10.48550/arXiv.1409.1556.
14. Hou, L., Samaras, D., Kurc, T.M., Gao, Y., Davis, J.E., and Saltz, J.H. (2016). Patch-based Convolutional Neural Network for Whole Slide Tissue Image Classification. Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. *2016*, 2424–2433. https://doi.org/10.1109/CVPR.2016.266.
15. 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. https://doi.org/10.1038/s41591-018-0177-5.
16. Wei, J.W., Tafe, L.J., Linnik, Y.A., Vaickus, L.J., Tomita, N., and Hassanpour, S. (2019). Pathologist-level classification of histologic patterns on resected lung adenocarcinoma slides with deep neural networks. Sci. Rep. *9*, 3358. https://doi.org/10.1038/s41598-019-40041-7.
17. Su, Z., Tavolara, T.E., Carreno-Galeano, G., Lee, S.J., Gurcan, M.N., and Niazi, M.K.K. (2022). Attention2majority: Weak multiple instance learning for regenerative kidney grading on whole slide images. Med. Image Anal. *79*, 102462.
18. 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.
19. Ilse, M., Tomczak, J.M., and Welling, M. (2018). Attention-based Deep Multiple Instance Learning. Preprint at ArXiv. https://doi.org/10.48550/arXiv.1802.04712.
20. Lu, M.Y., Chen, T.Y., Williamson, D.F.K., Zhao, M., Shady, M., Lipkova, J., and Mahmood, F. (2021). AI-based pathology predicts origins for cancers of unknown primary. Nature *594*, 106–110. https://doi.org/10.1038/s41586-021-03512-4.
21. 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. https://doi.org/10.1038/s41591-019-0462-y.
22. Lou, J., Xu, J., Zhang, Y., Sun, Y., Fang, A., Liu, J., Mur, L.A.J., and Ji, B. (2022). PPsNet: An improved deep learning model for microsatellite instability high prediction in colorectal cancer from whole slide images. Comput. Methods Programs Biomed. *225*, 107095. https://doi.org/10.1016/j.cmpb.2022.107095.
23. Echle, A., Grabsch, H.I., Quirke, P., van den Brandt, P.A., West, N.P., Hutchins, G.G.A., Heij, L.R., Tan, X., Richman, S.D., Krause, J., et al. (2020). Clinical-Grade Detection of Microsatellite Instability in Colorectal Tumors by Deep Learning. Gastroenterology *159*, 1406–1416.e11. https://doi.org/10.1053/j.gastro.2020.06.021.
24. 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. https://doi.org/10.1016/S1470-2045(20)30535-0.
25. Wang, T., Lu, W., Yang, F., Liu, L., Dong, Z., Tang, W., Chang, J., Huan, W., Huang, K., and Yao, J. (2020). Microsatellite Instability Prediction of Uterine Corpus Endometrial Carcinoma Based on H&E Histology Whole-Slide Imaging (IEEE), pp. 1289–1292.
26. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. Adv. Neural Inf. Process. Syst. *30*, 15–26.
27. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., and Gelly, S. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. Preprint at arXiv. https://doi.org/10.48550/arXiv.2010.11929.
28. Kotei, E., and Thirunavukarasu, R. (2023). A Systematic Review of Transformer-Based Pre-Trained Language Models through Self-Supervised Learning. Information *14*, 187.
29. Gao, X., Qian, Y., and Gao, A. (2021). Covid-vit: Classification of covid-19 from ct chest images based on vision transformer models. Preprint at arXiv. https://doi.org/10.48550/arXiv.2107.01682.
30. Karimi, D., Vasylechko, S.D., and Gholipour, A. (2021). Convolution-free Medical Image Segmentation Using Transformers (Springer), pp. 78–88.
31. Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., Lu, L., Yuille, A.L., and Zhou, Y. (2021). Transunet: Transformers make strong encoders for medical image segmentation. Preprint at arXiv. https://doi.org/10.48550/arXiv.2102.04306.
32. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., and Askell, A. (2020). Language models are few-shot learners. Adv. Neural Inf. Process. Syst. *33*, 1877–1901.
33. Zellers, R., Holtzman, A., Rashkin, H., Bisk, Y., Farhadi, A., Roesner, F., and Choi, Y. (2019). Defending against neural fake news. Adv. Neural Inf. Process. Syst. *32*, 9054–9065.
34. Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., and Joulin, A. (2020). Unsupervised learning of visual features by contrasting cluster assignments. Adv. Neural Inf. Process. Syst. *33*, 9912–9924.
35. Bulten, W., Kartasalo, K., Chen, P.H.C., Ström, P., Pinckaers, H., Nagpal, K., Cai, Y., Steiner, D.F., van Boven, H., Vink, R., et al. (2022). Artificial intelligence for diagnosis and Gleason grading of prostate cancer: the PANDA challenge. Nat. Med. *28*, 154–163. https://doi.org/10.1038/s41591-021-01620-2.
36. Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. Preprint at arXiv. https://doi.org/10.48550/arXiv.1810.04805.
37. Maron, O., and Lozano-Pérez, T. (1997). A framework for multiple-instance learning. Adv. Neural Inf. Process. Syst. *10*, 570–576.
38. Chévez-Barrios, P. (2021). Pathology of Uveal Melanoma. In Uveal Melanoma: Biology and Management, E.H. Bernicker, ed. (Springer International Publishing), pp. 37–51. https://doi.org/10.1007/978-3-030-78117-0_4.
39. Shah, J.P. (2015). Thyroid carcinoma: epidemiology, histology, and diagnosis. Clin. Adv. Hematol. Oncol. *13*, 3–6.
40. Magi-Galluzzi, C. (2018). Prostate cancer: diagnostic criteria and role of immunohistochemistry. Mod. Pathol. *31*, 12–21.
41. Diebold, J., Anderson, J.R., Armitage, J.O., Connors, J.M., Maclennan, K.A., Müller-Hermelink, H.K., Nathwani, B.N., Ullrich, F., and Weisenburger, D.D. (2002). Diffuse large B-cell lymphoma: a clinicopathologic analysis of 444 cases classified according to the updated Kiel classification. Leuk. Lymphoma *43*, 97–104.
42. Addis, B., and Roche, H. (2009). Problems in mesothelioma diagnosis. Histopathology *54*, 55–68.
43. Ho, J., Kalchbrenner, N., Weissenborn, D., and Salimans, T. (2019). Axial attention in multidimensional transformers. Preprint at arXiv. https://doi.org/10.48550/arXiv.1912.12180.
44. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., and Polosukhin, I. (2017). Attention Is All You Need, pp. 5998–6008.
45. Ba, J.L., Kiros, J.R., and Hinton, G.E. (2016). Layer normalization. Preprint at arXiv. https://doi.org/10.48550/arXiv.1607.06450.
46. Zhang, Z., and Sabuncu, M. (2018). Generalized cross entropy loss for training deep neural networks with noisy labels. Adv. Neural Inf. Process. Syst. *31*, 11–25.
47. Sing, T., Sander, O., Beerenwinkel, N., and Lengauer, T. (2005). ROCR: visualizing classifier performance in R. Bioinformatics *21*, 3940–3941. https://doi.org/10.1093/bioinformatics/bti623.
48. Clopper, C.J., and Pearson, E.S. (1934). The use of confidence or fiducial limits illustrated in the case of the binomial. Biometrika *26*, 404–413.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
Raw and analyzed data	https://portal.gdc.cancer.gov	TCGA
Raw and analyzed data	https://cancerimagingarchive.net/datascope/cptac	CPTAC
Raw and analyzed data	https://www.kaggle.com/c/prostate-cancer-grade-assessment/data	PANDA
Software and algorithms		
CLAM	(Lu et al.[18])	https://github.com/mahmoodlab/CLAM
TOAD	(Lu et al.[20])	https://github.com/mahmoodlab/TOAD

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and materials should be directed to and will be fulfilled by the lead contact, Xiangchun Li (lixiangchun2014@foxmail.com).

Materials availability

This study did not generate new unique reagents.

Data and code availability

- All datasets were downloaded from public databases. The source list of these datasets was provided in the key resources table. Source code is available at https://github.com/deeplearningplus/WIT.
- 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

WSI datasets

We collected a total number of 22,457 WSIs from The Cancer Genome Atlas (TCGA dataset, n = 11,623), The Clinical Proteomic Tumor Analysis Consortium (CPTAC dataset, n = 5,052) and PANDA (PANDA dataset, n = 5,782).

TCGA dataset

The TCGA dataset covers 32 cancer types: BRCA, KIRC, THCA, UCEC, LGG, LUSC, LUAD, HNSC, COAD, SKCM, PRAD, STAD, BLCA, GBM, LIHC, KIHC, CESC, SARC, PAAD, PCPG, READ, ESCA, TGCT, THYM, KICH, OV, UVM, MESO, UCS, ACC, DLBC and CHOL. The formalin-fixed paraffin embedded (FFPE) hematoxylin and eosin (H&E) stained WSIs are used. The details are in Table S1.

CPTAC dataset

We collected a total of 11,623 WSIs from the Cancer Imaging Archive CPTAC Pathology Portal. The collected projects consisted of CPTAC-LUAD, CPTAC-LSCC, CPTAC-SAR, CPTAC-UCEC, CPTAC-UCEC, CPTAC-CCRCC, CPTAC-PDA, CPTAC-HNSCC, CPTAC-SAR and CPTAC-CM (Table S8). The FFPE, H&E stained WSIs from normal donors and cancer patients are used.

The PANDA dataset

This dataset consists of 5,782 slides from prostate cancer patients and non-cancer individuals subjected to needle biopsies. There are 2,891 non-cancer biopsy WSIs. We randomly sampled 5,782 cancer biopsy WSIs to mitigate class imbalance cancer and non-cancer slides.

METHOD DETAILS

Whole-slide image (WSI) preprocessing

The slide image was segmented for the tissue regions using the CLAM Python package. We used ×20 magnification. We cropped the WSI into 256 × 256 patches within the segmented tissue regions and flattened them into an array. We extracted a feature of 1024 dimensions for these image patches from the second residual layer of pretrained ResNet50 model[12,34] on ImageNet dataset. The extracted features of image patches from a WSI were saved to disk file.

WIT architecture

WIT consists of an embedding layer and a transformer encoder followed by a softmax layer.

Embedding layer

This layer takes as input the elementwise summation of image patch features and position embeddings of the flattened image patches. We used the pretrained ResNet50 model[12,26] as the feature extractor for image patches.

The transformer encoder

The encoder has two components: a multi-headed self-attention and a position-wise feedforward neural network.

The i^{th} self-attention head is formulated as[26]:

$$Attention_i(Q_i, K_i, V_i) = softmax\left(\frac{Q_i K_i^T}{\sqrt{d_k}}\right)V_i$$

The input embeddings outputted from the embedding layer are projected to three matrices: query (Q_i), key (K_i) and value (V_i). d_k is the dimension of the query and it is used as scaling factor to mitigate the extreme small gradient.[44]

The multi-headed self-attention is the concatenation of multiple self-attention heads, allowing for the transformer attending to information in different feature representation subspaces. Multi-headed self-attention is formulated as[44]:

$$Multi - Head - Attention(Q, K, V) = Concat(Attention_1, \ldots, Attention_h)W^O$$

where $W^O \in \mathbb{R}^{hd_v \times d_{model}}$ denotes the learned projection matrix.

The position-wise feedforward neural network (*FFN*) consists of two linear layers with *ReLu* activation in-between:

$$FFN(x) = max(0, xW_1 + b_1)W_2 + b_2$$

where W_1 and W_2 are weight matrices and b_1 and b_2 are the bias.

Layer-wise normalization[45] is used in the front and rear of *FFN*. Residual connection[12] is applied to improve information flow.

Model training

The WSIs are random sampled and trained using WIT for 100 epochs. The weights and bias parameters of the model are initialized randomly, and the ground-truth label is slide-level labels. We used the cross-entropy loss[46] as the objective function in classification. The model parameters are updated via the *AdamW* optimizer with an initial learning rate of 2×10^{-5}, weight decay of 1×10^{-5}. WIT was trained with *PyTorch* (version 1.12.0) and *transformers* (version 4.21.1) on NVIDIA DGX A100.

Different WIT models

We evaluated four WIT models with different parameters by varying the hidden size: WIT-1Mb, WIT-2Mb, WIT-5Mb and WIT-17Mb. Details of these models are provided in Table S2.

Baseline method

We used attention-based MIL[18–20] implemented in TOAD[20] as baseline method. Attention-based MILs are widely used in computational pathology studies. It takes a WSI as a bag and image patches on that WSI as instances. It uses attention-based pooling to aggregate the features of all image patches to obtain slide-level feature representations.

Let $H = \{h_1, \ldots, h_k\}$ be a bag of K instances, the MIL pooling is defined as[30]:

$$z = \sum_{k=1}^{K} a_k h_k$$

a_k is the attention score for the k^{th} instance, which is defined as[30]:

$$a_k = \frac{\exp\left\{w_k^T \tanh\left(Vh_k^T\right)\right\}}{\sum_{j=1}^{K} \exp\left\{w_j^T \tanh\left(Vh_j^T\right)\right\}}$$

where $\forall_{k=1,\ldots,K}$, and $V \in \mathbb{R}^{L \times M}$ are parameters. The tanh is used as activation function. The network module is trained to assign an attention score a_t for each patch[30]:

$$a_k = \frac{\exp\left\{w_k^T\left(\tanh\left(Vh_k^T\right)\odot sigm\left(Uh_k^T\right)\right)\right\}}{\sum_{j=1}^{K}\exp\left\{w_j^T\left(\tanh\left(Vh_j^T\right)\odot sigm\left(Uh_k^T\right)\right)\right\}}$$

where $U \in \mathbb{R}^{L \times M}$ are parameters, $\odot$ is an element-wise multiplication and sigm(.) is sigmoid non-linearity.

Visualization of attention map

For a given self-attention head, let α is the self-attention matrix; $\alpha_{i,j}$ is the attention weight between the i^{th} and j^{th}. The attention score of the i^{th} patch with slide-level representation measures the contribution of the i^{th} patch on classification. *CLS* stands for a slide-level representation where we added at the start of flattened feature array of image patches for each WSI, which is used for classification during training. The self-attention is obtained via:

$$Softmax\left(\frac{Q_i \times K_i^T}{\sqrt{d_k}}\right)$$

Assumed there are *K* patches in a WSI, the first row of each self-attention matrix (denoted as α_0) quantifies the influence of each patch on classification. α_0 is converted to normalized percentile scores and scaled to the interval of [0, 1] as proposed in CLAM.[18] The normalized attention scores were converted to RGB colors using a disperse colourmap values and displayed on the spatial regions in the slide with high attention displayed in red and low attention in deep purple using Matlibplot (version 3.5.2). We tiled the WSI into 256 × 256 patches using a overlap of 0.80 to create more fine-grained heatmaps. Gaussian blur is used to smooth uneven pixel values in a heatmap image using OpenCV (version 4.7.0). We use the code of CLAM Python package for attention map visualization.[18] We used diverging color scheme (i.e., seismic palette in python matplotlib package) to represent the attention scores and overlay them onto the WSI image. The redder the higher probability of that region to be cancer, whereas the bluer the high probability of that region to be non-cancer.

QUANTIFICATION AND STATISTICAL ANALYSIS

Model evaluation

We used area under the receiver operating cureve (AUROC), accuracy, precision (also known as positive predictive value), recall rate, negative predictive value (NPV) and F1 score to assess the perfomance of WIT. Precision is the ratio of true positives to total predicted positives. Recall rate is the ratio of true positives to total actual positives. We reported the top-*K* accuracy for *K* = 1,2,3 on localization of 32 cancer types. NPV is defined as the number of true negatives divided by the number of samples predicted to be negative. F1-score is the harmonic mean of precision and recall rate.

Statistical and software

We conducted our experiment with Python (version 3.8.10), OpenSlide (version 1.2.0), Pillow (version 9.1.1), R (version 4.2.1), ggplot2 (version 3.3.6), ROCR (version 1.0.11), multiROC (version 1.1.1) and PROC[47] (version 1.18.0). The visualization of precision-recall curve (PRC) and calculation of area under PRC were performed with ROCR. Calculation of micro-averaged AUROC was performed with multiROC. Calculation of AUROC was performed with PROC.[47] The 95% confidence intervals of the AUROC were calculated using DeLong's methods implemented in pROC. The calculation of 95% confidence intervals for accuracy, sensitivity, specificity, precision, negative predictive value and F1 score with Clopper-Pearson method.[48]

Additional resources

This study did not generate additional data.