PAI-WSIT: a Comprehensive Curated Resource for Cancerous Pathology With Deep Learning

Changjiang Zhou
China Pharmaceutical University

Xiaobing Feng
China Pharmaceutical University

Yi Jin
China Pharmaceutical University

Harvest F. Gu
China Pharmaceutical University

Youcai Zhao
Nanjing First Hospital

Xiaodong Teng
Zhejiang University School of Medicine First Affiliated Hospital

Lingchuan Guo
First Affiliated Hospital of Soochow University

Jiatong Ji
China Pharmaceutical University

Shuopeng Jia
China Pharmaceutical University

Yan Xing
China Pharmaceutical University

Xiangshan Fan
Nanjing University Medical School Affiliated Nanjing Drum Tower Hospital

Jun Liao (✉ liaojun@cpu.edu.cn)
China Pharmaceutical University https://orcid.org/0000-0003-0617-5840

Research Article

Keywords: Artificial intelligence, Digital pathology, Database, Whole-slide images, Annotations

Posted Date: May 11th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-495066/v1

PAI-WSIT: a comprehensive curated resource for cancerous pathology with deep learning

Changjiang Zhou [a,#], Xiaobing Feng [b,#], Yi Jin [a], Harvest F. Gu [b], Youcai Zhao [c], Xiaodong Teng [d], Lingchuan Guo [e], Jiatong Ji [b], Shuopeng Jia [b], Yan Xing [a], Xiangshan Fan [f,]* & Jun Liao [a,]*

[a] *School of Science, China Pharmaceutical University, Nanjing, China*

[b] *School of Basic Medicine and Clinical Pharmacy, China Pharmaceutical University, Nanjing, China*

[c] *Department of Pathology, Nanjing First Hospital, Nanjing, China*

[d] *Department of Pathology, the First Affiliated Hospital, College of Medicine, Zhejiang University, Hangzhou, China*

[e] *Department of Pathology, the First Affiliated Hospital of Soochow University, Soochow, China*

[f] *Department of Pathology, Nanjing Drum Tower Hospital, Nanjing, China*

* Corresponding authors.

E-mail addresses: liaojun@cpu.edu.cn (J. Liao); fxs23@163.com (X.S. Fan).

[#] Changjiang Zhou and Xiaobing Feng contributed equally to this work.

Abstract

Background

The possibility of digitizing whole-slide images (WSI) of tissue has led to the advent of artificial intelligence (AI) in digital pathology. Advances in precision oncology have resulted in an increasing demand for predictive assays that enable mining of subvisual morphometric phenotypes and might improve patient care ultimately. Hence, a pathologist-annotated and artificial intelligence-empowered platform for integration and analysis of WSI data and molecular detection data in tumors was established, called PAI-WSIT (http://www.paiwsit.com).

Methods

The standardized data collection process was used for data collection in PAI-WSIT, while a multifunctional annotation tool was developed and a user-friendly search engine and web interface were integrated for the database access. Furthermore, deep learning frameworks were applied in two tasks to detect malignant regions and classify phenotypic subtypes in colorectal cancers (CRCs), respectively.

Results

PAI-WSIT recorded 8633 WSIs of 1772 tumor cases, of which CRC from four regional hospitals in China and The Cancer Genome Atlas (TCGA) were the main ones, as well as cancers in breast, lung, prostate, bladder, and kidneys from two Chinese hospitals. A total of 1298 WSIs with high-quality annotations were evaluated by a panel of 8 pathologists. Gene detection reports of 582 tumor cases were collected. Clinical information of all tumor cases was documented. Besides, we reached overall accuracy of 0.933 in WSI classification for malignant region detection of CRC, and aera under the curves (AUC) of 0.719 on colorectal subtype dataset.

Conclusions

Collectively, the annotation function, data integration and AI function analysis of PAI-WSIT provide support for AI-assisted tumor diagnosis, all of which have provided a comprehensive curation of carcinomas pathology.

Keywords Artificial intelligence, Digital pathology, Database, Whole-slide images, Annotations

Background

Pathology is considered as the “gold standard” of diagnostic medicine and directly related to the subsequent treatment. Instead of conventional microscopy, digital pathology plays a crucial role in modern clinical practice in recent years due to its advances in computing power, fast networks, and large storage [1, 2]. With the digital slide scanners, whole slide images (WSIs) become a great source of information and complexity in pathology because of their large size (commonly at a resolution of 100k×100k) [3]. Furthermore, artificial intelligence (AI), particularly deep learning, is bringing a paradigm shift to many breakthroughs in image classification, object detection and segmentation [4]. AI algorithms have the potential for developing an unifying approach in digital pathology [5], including segmentation and classification of various regions in WSIs [6, 7], detection of tumor proliferation [8] and cancer metastases [9], mitosis detection [10], as well as prediction of patient prognosis [11, 12]. Several pathology datasets with AI, for example, CAMELYON16 challenge for breast cancer metastasis detection [13], and MoNuSeg2018 challenge for multi-organ nuclei segmentation [14] have demonstrated that they are highly useful for facilitating disease studies and diagnosis tool developments in AI [15].

Rather than being a single, uniform disease type, accumulating evidence suggests that cancer comprises a group of molecularly heterogeneous diseases that are characterized by a range of genomic and epigenomic alterations. Pathology diagnosis depends not only on WSI analysis but also on several other sources of data that need to be included coming from omics, clinical records, and patient medical information. For example, immunohistochemistry (IHC) staining has a profound role in diagnosis by helping doctors to determine the biological characteristics of a wide variety of tumors, to make a prognosis, and to select appropriate systemic therapies for patients with cancer [16]. As the use of genomic technologies spreading, DNA-level and transcriptional-level features obtained from tissues will be evaluated for their utility in creating molecular subtypes of cancer to predict future disease behavior and treatment response [17]. For example, the Pan-Cancer Atlas reclassifies human tumor types based on molecular similarity [18], indicating that the cell of origin influences but does not fully determine tumor classification, which informs future clinical trial design and interpretation.

In 2015, The CRC Subtyping Consortium (CRCSC) was formed to show marked interconnectivity between six independent classification systems coalescing into four consensus molecular subtypes (CMSs) with distinguishing features: CMS1 (microsatellite instability immune); CMS2 (canonical); CMS3 (metabolic); and CMS4 (mesenchymal) [19]. Most clinicians consider the CMS groups would be taken forward into routine clinical practice with the aim of prognostic value [20]. AI is essential to recognize the phenotypic features of CRC, discover more complicated or subtle connections than a human would and help pathologists make the best clinical decisions for patients.

Thus, we have established an artificial intelligence multicenter platform for integration and analysis in (www.paiwsit.com). The standardized data collection process was used for data collection, a multifunctional annotation tool was developed and a user-friendly search engine and web interface were integrated for database access. To facilitate data analysis and interpretation based on PAI-WSIT, a systematic analytical framework was also proposed. A deep learning model ResNet for malignant regions detection in colorectal WSI with slides and annotations were provided, meanwhile, we actively explored whether it is possible to classify four CMSs of CRC only using WSI via deep learning, which lay the foundation for the deployment of computational decision support systems for CRC in clinical practice.

Methods

Data collection and curation

As a multicenter platform, we collected pathological data of cancer patients from four hospitals (Nanjing First Hospital, the First Affiliated Hospital of Zhejiang University, the First Affiliated

Hospital of Soochow University and Nanjing Drum Tower Hospital) and TCGA. With cancer patients as the basic elements, PAI-WSIT recorded physiological information, clinical and pathological reports, WSIs and gene mutation data from detection reports.

WSIs with hematoxylin-eosin (H&E) and immunohistochemistry staining were stored. Slides were scanned by Hamatsu NanoZoomer C9600-12 at a magnification of 40x (image resolution: 0.225 μm/pixel). The biomarker mutation data were generated based upon the target region probe capture technology and sequenced by next-generation sequencing (NGS) technology using Illumina platform and the detection of plasma and tissue samples from cancer patients. Clinical data such as physiological parameters and pathological diagnosis information of patients were mostly in unstructured free-text reports. With these metadata, researchers were able to annotate WSIs and also train deep learning algorithms for more tasks such as lymphoma type classification. A detailed listing of all the fields of metadata and their descriptions (Additional file1: Table S1) are provided, meanwhile.

Data annotation

Lack of high-quality annotation is one of the biggest challenges for AI in digital pathology [21]. In PAI-WSIT, we developed the annotation tools, including smooth curve tool, polygon tool and other shaped tools. Using these tools, the users could draw shapes on the WSIs to highlight various kinds of cells or regions on the slides and thereby enable them to annotate the WSIs in higher accuracy with less effort. For each user, all annotations for a given WSI were grouped in one annotation tree (**Fig. 1(a)**). More importantly, PAI-WSIT supported the storing annotations from multiple pathologists for the same WSI. Pathologists or other platform users could view annotations drawn by other pathologists. Multi-user annotations provided the possibility and convenience to compare different annotations and to adopt a flexible data processing strategy for deep learning case studies.

We invited several experienced pathologists annotated colorectal cancer WSIs at the pixel levels by using the annotation tools. To control the quality of annotations, each slide was annotated by two pathologists. An independent expert of pathology further examined the quality of annotations, and merged the annotations for each slide into a definitive one. The data were then stored on a dedicated server cluster and a web interface provided the convenient access of data, which could be referenced by the users of PAI-WSIT. The annotation protocol was shown in **Fig. 1(b)**.

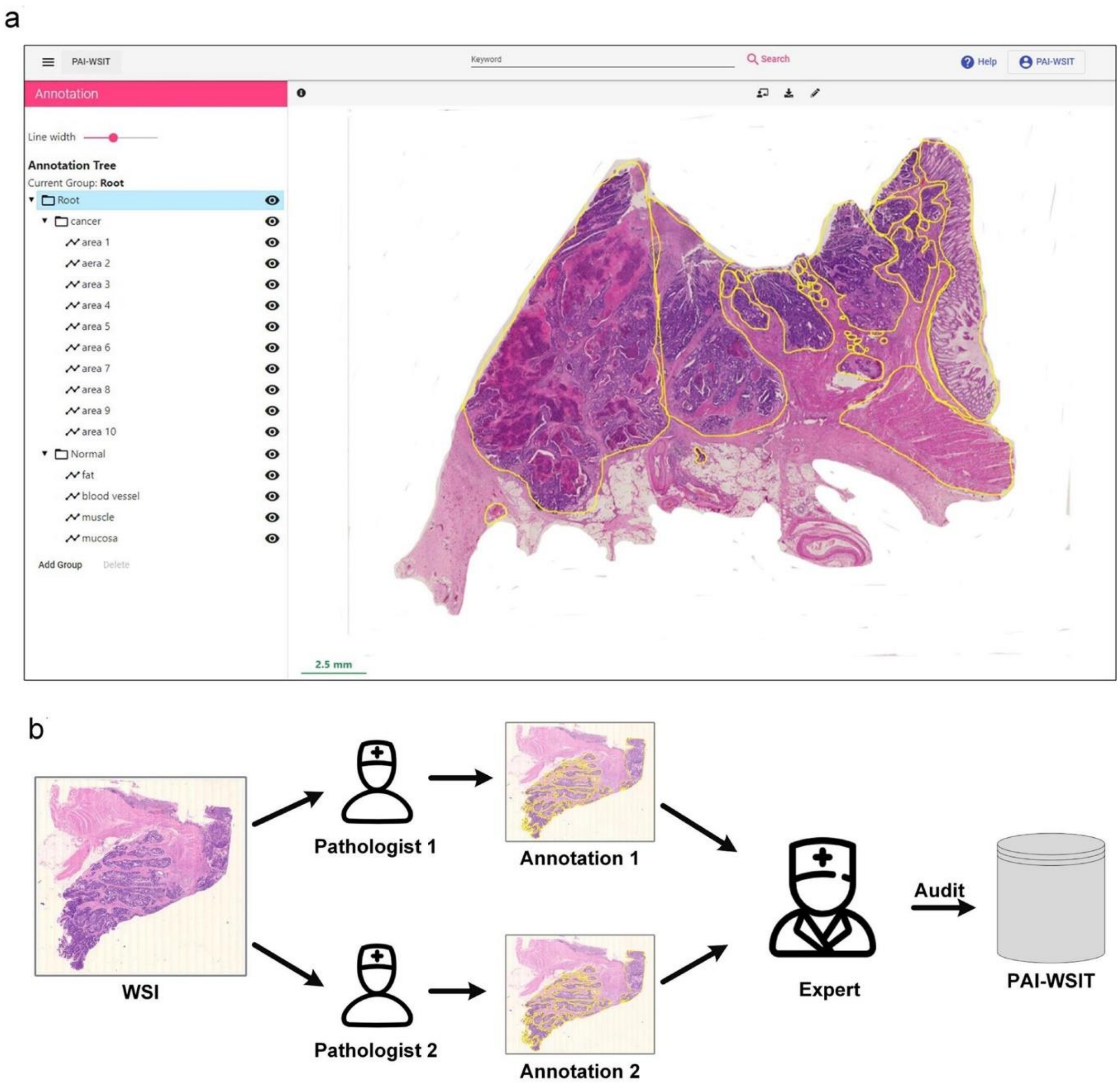

Fig. 1. Annotation tool in PAI-WSIT. (a) Annotation tree: the user draws two groups of regions in a WSI, one is cancer areas and the other is normal areas. (b) Annotation protocol: each annotation in PAI-WSIT was annotated by two pathologists and audited by an experienced expert. *WSI* whole slide image

Database architecture

The overall workflow of PAI-WSIT is summarized in **Fig. 2**. PAI-WSIT collected large pathological data of patients, such as basic information, pathological reports, WSIs and gene detection reports and also integrated annotation tools and deep learning models. Moreover, a number of annotations produced by experienced pathologists, which lay the foundation for the deployment of computational decision support systems in clinical practice.

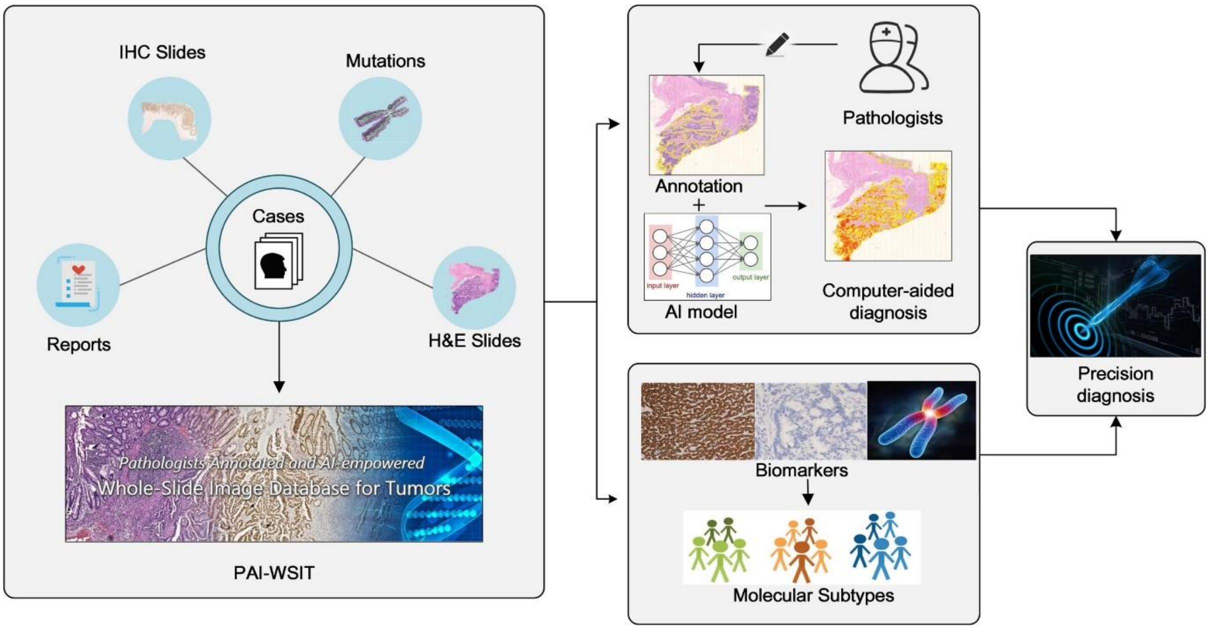

Fig. 2. The overall workflow of PAI-WSIT. PAI-WSIT contains basic information, pathological reports, WSIs and target biomarker results of cancer cases. The integration of annotation tools towards WSI makes AI in digital pathology more convenient. Biomarkers in DNA-level and transcriptional-level are helpful for molecular subtype research. *IHC* immunohistochemistry, *H&E* hematoxylin-eosin, *AI* artificial intelligence

A user-friendly search engine and web interface were incorporated for users to search and browse WSIs along with their metadata and annotations. Outstandingly, PAI-WSIT can visualize the result generated by AI analysis for colorectal WSIs and relevant studies. PAI-WSIT takes advantages of various storage technique, including distributed file system (CephFS [22]), a relational database (MySQL) and a document database (MongoDB [23]), as well as a search engine (Elasticsearch [24]) providing functionalities of searching data synchronized from MongoDB database. For web browsing, the front page of the retrieval system was based on HTML. The web interface is a single page application built using Angular framework to provide a user-friendly experience as well as to support sophisticated features such as annotation management and various annotation tools.

Framework for malignant regions detection (in colorectal WSIs)

We trained a deep learning model ResNet for malignant region detection in colorectal WSI with slides and annotations from PAI-WSIT. Both training sets and test sets were H&E slides from Nanjing First Hospital and validation set were slides from Nanjing Drum Tower Hospital. The summary of patient information in this study and the distribution of WSIs are represented in Table 1.

Table 1 Summary of patient information and the distribution of WSIs

Data set	**Training set**	**Test set**	**Validation set**
Sex			
(MALE/FEMALE)	73/57		28/22
Age	62.5±12.1		64.2±13.2
Normal (patients/WSIs)	40/40	13/13	21/21
Cancer (patients/WSIs)	60/60	17/17	29/29

WSI whole slide image

Several efforts were put in the technical validation of PAI-WSIT. First, regions of interest (ROIs) are calculated for the requested WSI to avoid processing the white background of the WSI. Next, the images of the ROIs were cropped with 256*256 patches with pathologists' annotations (positive samples for cancer regions and negative samples for tumor regions) as that of in the original model input. There was an unbalance between positive and negative samples, so cancer samples were augmented by extracting the rectangle until the number of normal and cancer samples approximately equal finally.

ResNet utilizes shortcut connections to significantly reduce the difficulty of training, which achieves competitive performances in classification compared with other kinds of networks. Therefore, ResNet is used in our case. It starts a stochastic gradient descent (SGD) at a learning rate of 0.001. we chose cross-entropy as loss function:

$$H_{y'}(y) = \ -\sum_i y_{i'}\log(y_i) \tag{1}$$

Where y_i denotes prediction and y_i' denotes the ground truth. We generated probability maps using the trained model and drew heatmap for visualization, with which even can achieve the location of the tumor area. The whole process is shown in **Fig. 3**.

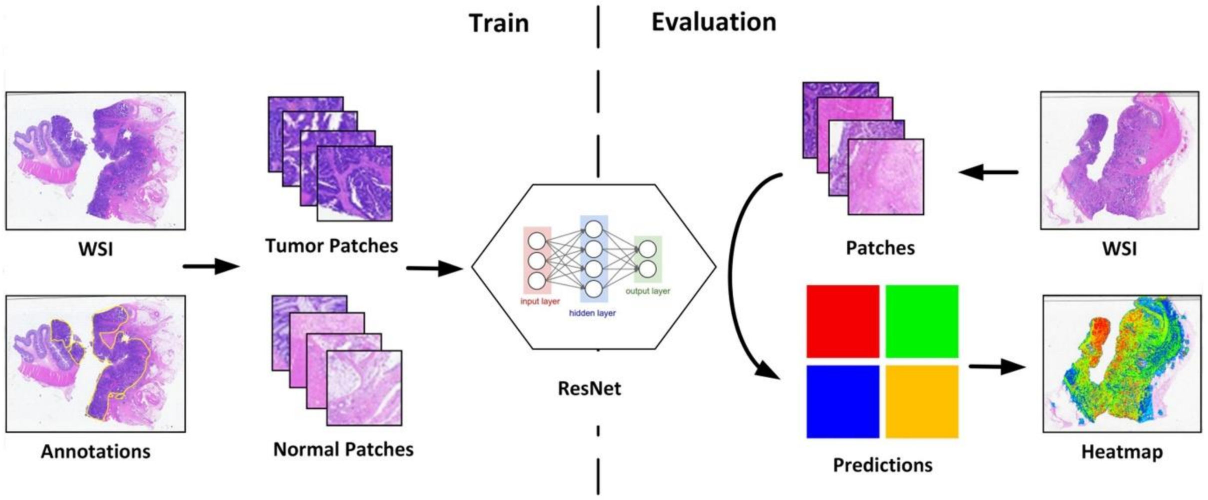

Fig. 3. AI workflow. Patches were extracted by WSIs in training set, and label as positive or negative samples by annotation. After training, WSIs in validation set were predicted and visualized as heatmaps. *WSI* whole slide image

AI in molecular subtypes of CRC

Molecular subtypes of CRC can be characterized by genomic and phenotypic features. Given that different subtypes have different outcomes, the ability to subtype tumors in the clinical practice would be highly favorable, enabling optimal treatment for individual patients [20]. In PAI-WSIT, the data of molecular biomarkers and IHC results from patients were recorded, which were helpful for researchers to explore the molecular subtypes of cancer.

Given that WSI contains a wealth of histological features about CRC, and would present a more readily translatable method for subtyping CRC tumors. Therefore, one of the purposes of this study was to investigate whether we can classify CMS of CRC using WSI by deep learning, since WSI is more easily acquired in the routine clinical pathology laboratory. The whole process is shown in **Fig. 4**.

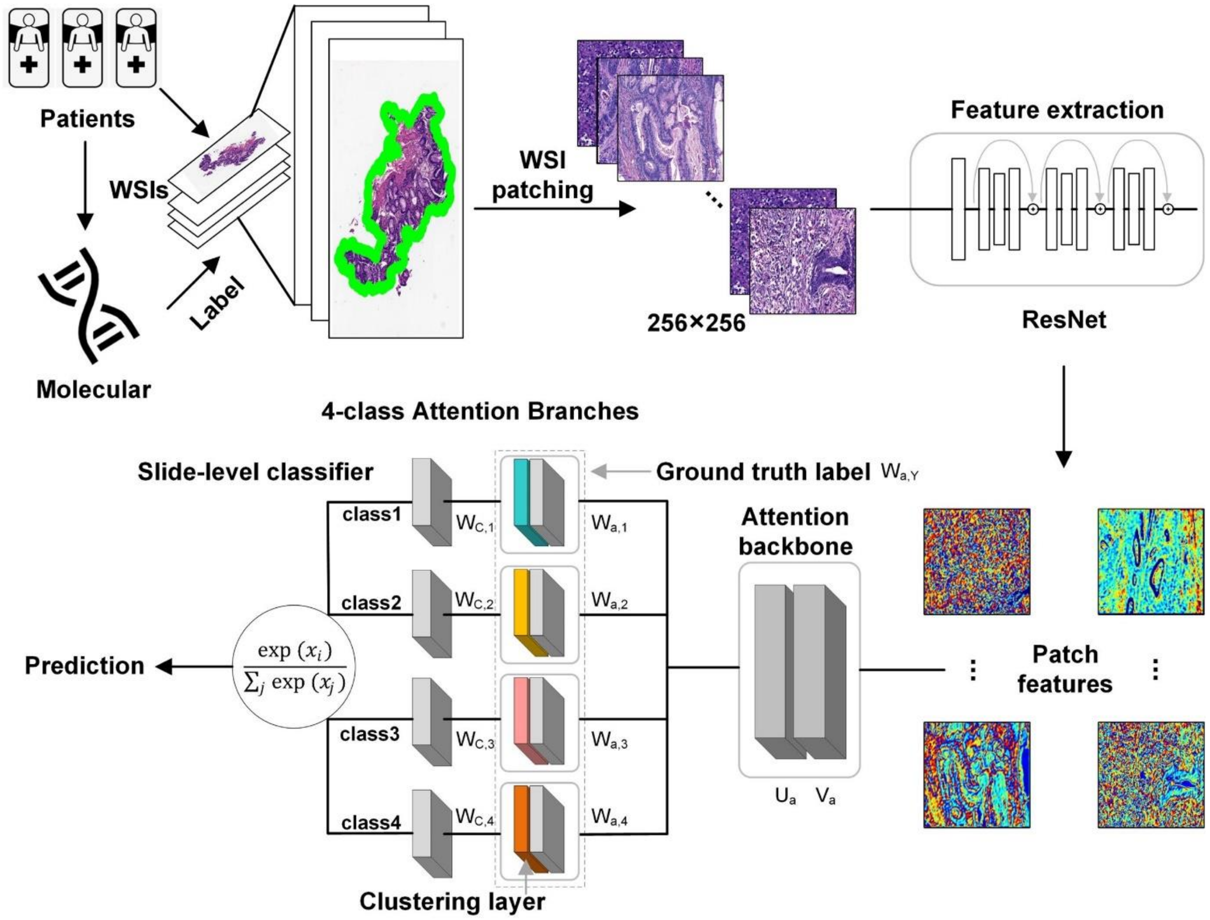

Fig. 4. Pipeline of CMS classification**.** The WSIs of CMS were labeled by gene expression data, all patches are extracted from tissue regions of a WSI. Patches are encoded once by a pretrained ResNet into a descriptive feature representation. Attention-pooling weighs patches and summarizes patch-level features into slide-level representations to make final prediction. *WSI* whole slide image

For each digitized slide, the corresponding label was lied on the gene expression data of each patient. After segmentation, 256×256 patches were generated to be encoded by a pretrained ResNet. Based on the clustering-constrained attention multiple instance learning model [25], attention-based learning was applied for aggregating patch-level features into slide-level representations for classification. Following features extraction, both training and inference can be carried out in the low-dimensional feature space. In instance-level clustering, for each of 4 classes, we place a clustering layer with 512 hidden units after the first fully-connected layer, $W1$. $W1 \in \mathbb{R}^{512\times1024}$ further compress each fixed 1024-dimensional patch-level representation z_k to a 512-dimensional vector $h_k = W_1 z_k^T$. In the 4-calss classification task, the first two layers of the attention network $U_a \in \mathbb{R}^{256\times512}$ and $V_a \in \mathbb{R}^{256\times512}$ collectively as the attention backbone shared by all classes, the attention network splits into 4 parallel attention branches $W_{a,1}$,…, $W_{a,4} \in \mathbb{R}^{1\times256}$. Similarly, 4 parallel independent classifiers, $W_{c,1}$,…, $W_{c,4}$ are built to score each class-specific slide-level representation. Accordingly, the attention score of the k^{th} patch for the 4 classes, denoted $a_{k,4}$, is given by Eq2 and the slide-level representation aggregated per the attention score distribution for the 4 classes, denoted $h_{slide,4} \in \mathbb{R}^{1\times512}$, is given by Eq3:

$$a_{k,4} = \quad (2)$$

$$\frac{exp\{\boldsymbol{W}_{a,4}(\tanh(V_a h_k^T)\odot sigm(U_a h_k^T))\}}{\sum_{j=1}^{N} exp\{\boldsymbol{W}_{a,4}(\tanh(V_a h_k^T)\odot sigm(U_a h_k^T))\}}$$

$$h_{slide,4} = \sum_{k=1}^{N} a_{k,4} h_k \quad (3)$$

The corresponding unnormalized slide-level score $s_{slide,4}$ is given via the classifier layer $\boldsymbol{W}_{c,4} \in \mathbb{R}^{1\times256}$ by $s_{slide,4} = \boldsymbol{W}_{c,4} h_{slide,4}^T$. For inference, the predicted probability distribution over each class is computed by applying a softmax function to the slide-level prediction scores.s_{slide} .

Our data set was derived from TCGA, including 391 cases, 731 WSIs. The CMS slides with annotated ROIs from gene expression data were cropped with 256×256 patches as input to the original model. Finally, we evaluated the slide-level classification performance using 10-fold monte carlo cross-validation. For each cross-validate fold, we randomly partitioned WSI dataset into a training set (80% of cases), a validation set (10% of cases) and test set (10% of cases). The distribution of WSIs is represented in Table 2.

Table 2 Distribution of WSIs Dataset

	CMS1	**CMS2**	**CMS3**	**CMS4**
Cases	63	176	57	95
WSIs	124	330	101	176
Patches	606520	1682444	509845	1174840

CMS consensus molecular subtyping, *WSI* whole slide image

Results

Data statistics in PAI-WSIT

PAI-WSIT collected 1772 cases, including colorectal cancer, breast cancer, prostate cancer, lung cancer, kidney cancer and bladder cancer from four hospitals in China and TCGA. The metadata of all cares such as basic information, pathological diagnosis reports were recorded. Furthermore, 8663 WSIs, including H&E slides and IHC slides and 3664 high-quality annotations generated by 8 experienced pathologists H&E slides, were also collected. Currently, PAI-WSIT occupies 7.67 terabytes of disk space of the storage cluster, and the summary and distribution of data are shown in **Fig. 5**. All information on the cases including basic information, pathological diagnosis reports, H&E slides, IHC slides and high-quality annotations, are available at www.paiwsit.com.

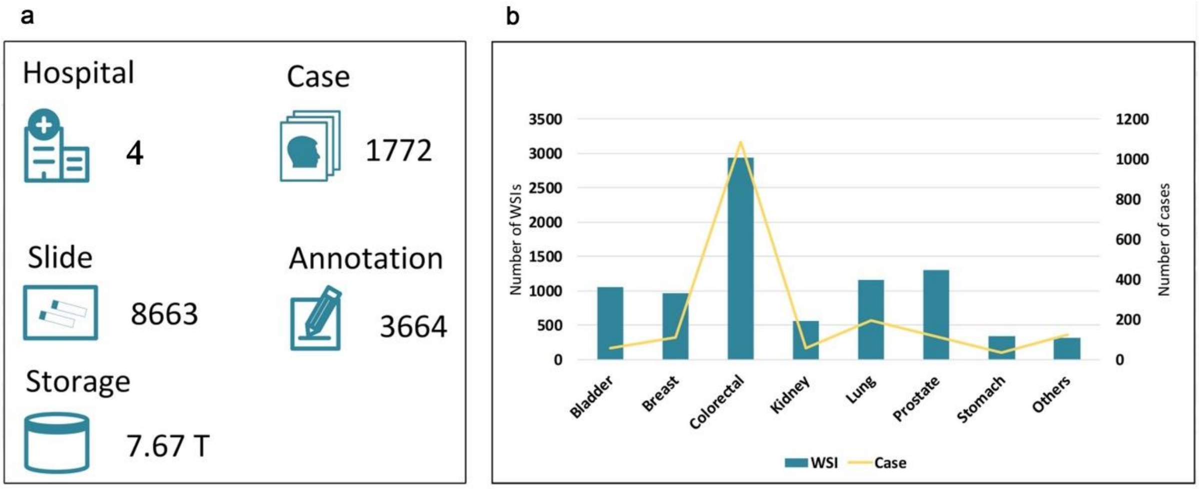

Fig. 5. An overview of the data PAI-WSIT has recorded. (a) Data summary in PAI-WSIT. (b) Cases and slides by major primary site in PAI-WSIT. *WSI* whole slide image

Web interfaces in PAI-WSIT

A user-friendly web interface and search engine were incorporated for researchers in PAI-WSIT (www.paiwsit.com), which allows you to store your digital pathology images into the cloud platform so that you can search and view the images in the browser. PAI-WSIT also provides a feature-rich online tool for annotating your digital slides. Also, you can view the annotations drawn by other pathologists. For colorectal cancer, PAI-WSIT leverages AI technology to detect the abnormal regions in whole slide images. Results can be achieved and visualized in the form of heatmaps. The usage documentary (Additional file2: Figure S1) of PAI-WSIT is provided in a detailed description. Moreover, several help documents, including account management, WSI collection management, patient case management, WSI viewer usage and search engine usage are listed on (http://www.paiwsit.com/help), which may instruct you how to use the platform.

Application for Binary Positive vs. Negative Classification (in colorectal WSIs)

The classification results of test and validation sets in patch-level and image-level are listed in Table 3. Benchmark metrics of prediction in patch-levels are better than the image-levels, especially in accuracy. Although all benchmark metrics in validation are little worse than them in testing set, the performance is satisfactory and suggests that the deep learning-based approaches can be broadly applied in histopathology image field to assist pathologists in dealing with clinical tasks. With the help of heatmap, it even can map the location of the tumor area, and the results are shown in **Fig. 6**.

Table 3 The results of ResNet in both testing and validation sets at patch-level and image-level

Dataset	Level	Accuracy	Recall	F1 score
Testing	Patch	0.967	0.947	0.951
	Image	0.933	0.941	0.941
Validation	Patch	0.953	0.978	0.939
	Image	0.920	0.966	0.933

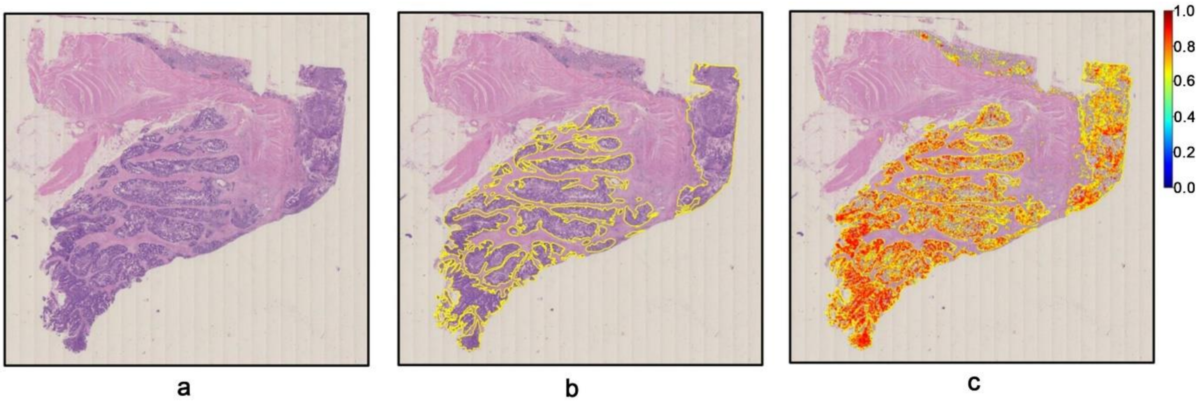

Fig. 6. Prediction of location of tumor regions in ResNet. (a) Raw WSI;(b) Annotation; (c) Prediction. For each annotated slide, the whole slide attention heatmap is generated by their tumor probability. *WSI* whole slide image

AI for Subtyping Problem (4-class CRC Subtyping)

We evaluated the slide-level classification performance using 10-fold monte carlo cross-validation. For each cross-validate fold, we randomly partitioned WSI dataset into a training set (80% of cases), a validation set (10% of cases) and test set (10% of cases). To investigate the dependency of the model's performance on the amount of training data available, for each cross-validated fold created, we sequentially sampled subsets of training data equal to 75%,50%,25%,10% of the total number of cases in the training set, while keeping the validation and test set the same.

We observed that the model is more confident in its correct predictions than in its incorrect predictions and becomes less and less confident as the size of the training set is reduced **(Fig.7(a-d))**. When testing on the test cohort, the 10-fold cross-validated model trained using

100% of the training set achieved an average one-vs-rest AUC (macro-averages) of 0.719 on colorectal subtype dataset **(Fig.7(e))**, which reflects the shortcomings of quality and number of the dataset.

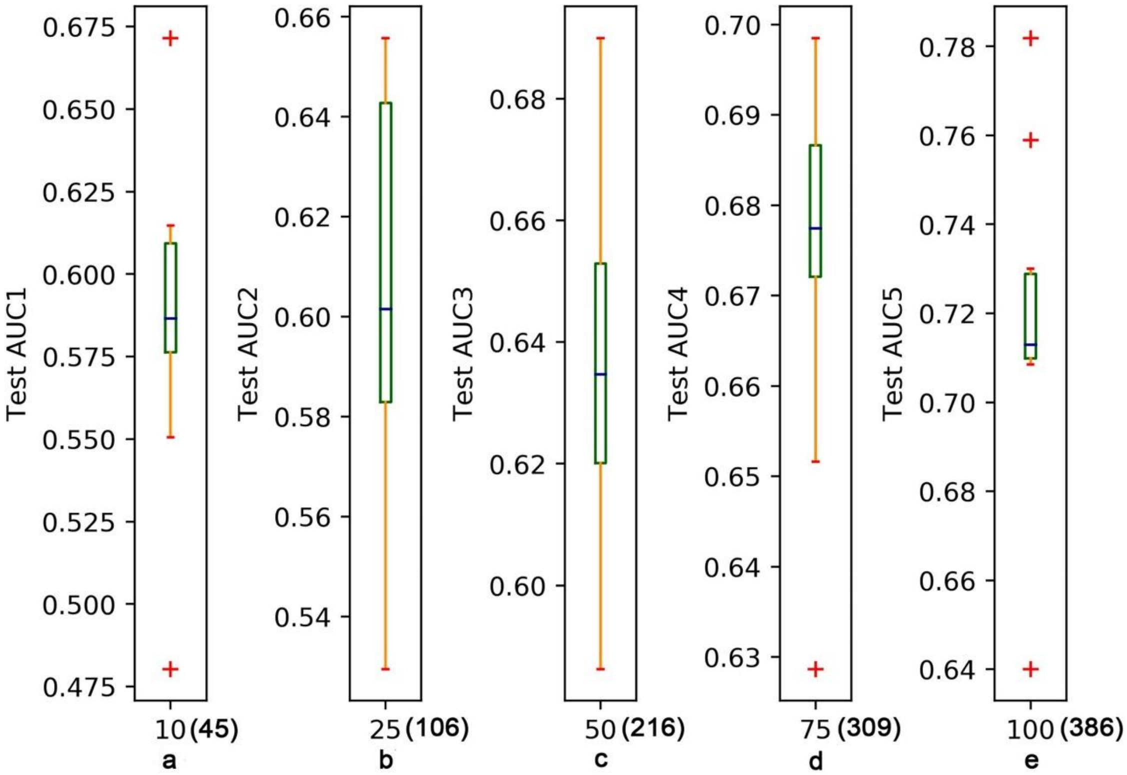

Fig. 7. AUC of test set. (a-e) were % of Training set used (Number of WSIs). *AUC* area under the curve, *WSI* whole slide image

Discussion

There is increasing evidence that AI will be essential to integrate disparate sources of information, discover more complicated or subtle connections than a human would and help pathologists make the best clinical decisions for patients [17, 26, 27]. To address the data integration and analysis for computational pathology mainly in colorectal cancer, we developed a pathologist-annotated for WSI database. We also proposed frameworks for binary task and subtyping problem with standard annotated WSIs. Therefore, the platform of PAI-WSIT may serve as an excellent source for both clinical and research settings in tumors precision diagnosis, and create the collaboration among clinical pathologists to further develop the computer-aided prognosis as well.

The aim of present study was to provide data for artificial intelligent researches to fuel the development of algorithms and tools in cancer research and clinical diagnosis. To achieve this goal, as a multicenter platform, PAI-WSIT collected pathological data of cancer patients from four regional hospitals in China and TCGA. With cancer patients as the basic elements, PAI-WSIT recorded physiological information, clinical and pathological reports, WSIs and gene mutation data from detection reports. More importantly, we integrated annotation tools and deep learning models for malignant regions detection and molecular subtype classification of CRCs.

Accurate detection of malignant regions in colorectal WSIs is a perfect combination between AI and histopathology, while phenotypic subtyping could be the first step toward precision medicine for CRC. Routine subtyping for CRC tumors could revolutionize the treatment of patients, with each subtype receiving a different therapeutic regime. For example, it is already known that the EGFR monoclonal antibody, cetuximab may not be useful for patients with the CMS1 or CMS3 subtypes [28]. Nevertheless, CMS1 appears to be the only subtype sensitive to Src family kinase inhibitors such as Dasatinib [29]. The study result has also been shown that CMS4 is resistant to chemotherapy [30]. Though the results of AI models in the performance of CRC molecular subtypes were unsatisfactory, it has laid the foundation for future work improvement.

Annotating the 1298 colorectal slides with high-quality is a valuable effort, being comparable to an earlier breast cancer dataset [13], though those do not contain functionality for reading annotations or storing image analysis results. Given that the sizes of datasets in recent studies focusing on a specific disease are exceeding the 10000 WSI mark [31] and digital pathology is moving towards a more epidemiological scale [32-34], we are working with hospitals and pathologists to further improve and popularize PAI-WSIT with more WSIs and annotations which makes it a valuable resource in computational pathology in the very near future.

There are two limitations of this study. First, the prediction results of AI in the molecular subtypes of CRCs are slightly unsatisfactory. In future work, artificial subjective bias caused by slide's quality should be avoided, post-processing methods should be improved, and better models such as U-NET [35] should be tried. Second, and more importantly, it should be more than just WSI analysis to increase physicians' understanding of AI outcomes. Thus, we will integrate other omics data, including proteomics and genomics, in future research.

Conclusions

PAI-WSIT (https://www.paiwsit.com) contains histopathology, immunohistochemistry, gene mutations and clinical data of tumor patients, forming pathological big data. The platform's annotation function and deep learning models provide support for AI-assisted tumor diagnosis, and can serve as a valuable resource for use by the community of researchers in the WSI analysis and tissue diagnostics domains.

Supplementary information

Supplementary data are available at Journal of Translational Medicine online.

Abbreviations

AI: Artificial intelligence; WSI: Whole slide image; CMS: Consensus molecular subtyping; CRCs: Colorectal cancers; IHC: Immunohistochemistry; H&E: Hematoxylin-eosin; NGS: Next-generation sequencing; ROIs: Regions of interest; SGD: Stochastic gradient descent; CRCSC: The Colorectal Cancer Subtyping Consortium; TCGA: The Cancer Genome Atlas; AUC: area under the curve.

Acknowledgments

This study was supported by the High-Performance Computing Center, China Pharmaceutical University.

Authors' contributions

C.Z. carried out the main research. C.Z. and X.F. wrote the manuscript. Y.J. designed the Web application. H.F.G. polished the manuscript. X.F., Y.Z., L.G. and X.T. collected and annotated the data. J.J., S.J., Y.X. and X.F. interpreted the data. J.L. contributed to method development and reviewed the manuscript.

Funding

This work was financially supported by the Double-Class University project (CPU2018GY19), the National Natural Science Foundation of China (81473274), and the Postgraduate Research & Practice Innovation Program of Jiangsu Province (KYCX_19_0667).

Availability of data and materials

The curated patience cases along with WSIs were stored in a MySQL database cluster and a distributed file system deployed on a group of on-premises servers. A feature-rich web application was developed to provide public access to the data. The backend of the application is developed using ASP.NET Core, while the frontend is developed using Angular, which communicates with the backend using REST APIs. To help researches fetch patient information (including WSIs and their annotations) programmatically from the platform, a library written in C# is available on our GitHub repository (https://github.com/yigolden/paiwsit-client-csharp).

Ethics approval and consent to participate

The study was approved by the Ethical Committee of Nanjing First Hospital, the First Affiliated Hospital of Zhejiang University, the First Affiliated Hospital of Soochow University and Nanjing Drum Tower Hospital. Due to the retrospective nature of the study, written informed consent for participation in the study was waived.

Consent for publication

All the authors in this paper consent to publication of the work.

Competing interests

The author declares that they have no competing interests.

References

1. Fuchs TJ, Buhmann JM: Computational pathology: challenges and promises for tissue analysis. Comput Med Imaging Graph 2011, 35:515-530.
2. Louis DN, Feldman M, Carter AB, Dighe AS, Pfeifer JD, Bry L, Almeida JS, Saltz J, Braun J, Tomaszewski JE, et al: Computational Pathology: A Path Ahead. Arch Pathol Lab Med 2016, 140:41-50.
3. Niazi MKK, Parwani AV, Gurcan MN: Digital pathology and artificial intelligence. Lancet Oncol 2019, 20:e253-e261.
4. LeCun Y, Bengio Y, Hinton G: Deep learning. Nature 2015, 521:436-444.
5. Janowczyk A, Madabhushi A: Deep learning for digital pathology image analysis: A comprehensive tutorial with selected use cases. J Pathol Inform 2016, 7:29.
6. Ehteshami Bejnordi B, Veta M, Johannes van Diest P, van Ginneken B, Karssemeijer N, Litjens G, van der Laak J, Hermsen M, Manson QF, Balkenhol M, et al: Diagnostic Assessment of Deep Learning Algorithms for Detection of Lymph Node Metastases in Women With Breast Cancer. Jama 2017, 318:2199-2210.
7. Khosravi P, Kazemi E, Imielinski M, Elemento O, Hajirasouliha I: Deep Convolutional Neural Networks

Enable Discrimination of Heterogeneous Digital Pathology Images. EBioMedicine 2018, 27:317-328.

8. Cardoso MJ, Arbel T, Carneiro G, Syeda-Mahmood T, Tavares J, Moradi M, Bradley A, Greenspan H, Papa J, Madabhushi A, et al: Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: Third International Workshop, DLMIA 2017, and 7th International Workshop, ML-CDS 2017, Held in Conjunction with MICCAI 2017, Québec City, QC, Canada, September 14, Proceedings. 2017.

9. Golden JA: Deep Learning Algorithms for Detection of Lymph Node Metastases From Breast Cancer: Helping Artificial Intelligence Be Seen. Jama 2017, 318:2184-2186.

10. Saha M, Chakraborty C, Racoceanu D: Efficient deep learning model for mitosis detection using breast histopathology images. Comput Med Imaging Graph 2018, 64:29-40.

11. Bychkov D, Linder N, Turkki R, Nordling S, Kovanen PE, Verrill C, Walliander M, Lundin M, Haglund C, Lundin J: Deep learning based tissue analysis predicts outcome in colorectal cancer. Sci Rep 2018, 8:3395.

12. Mobadersany P, Yousefi S, Amgad M, Gutman DA, Barnholtz-Sloan JS, Velázquez Vega JE, Brat DJ, Cooper LAD: Predicting cancer outcomes from histology and genomics using convolutional networks. Proc Natl Acad Sci U S A 2018, 115:E2970-e2979.

13. Litjens G, Bandi P, Ehteshami Bejnordi B, Geessink O, Balkenhol M, Bult P, Halilovic A, Hermsen M, van de Loo R, Vogels R, et al: 1399 H&E-stained sentinel lymph node sections of breast cancer patients: the CAMELYON dataset. Gigascience 2018, 7.

14. Kumar N, Verma R, Sharma S, Bhargava S, Vahadane A, Sethi A: A Dataset and a Technique for Generalized Nuclear Segmentation for Computational Pathology. IEEE Trans Med Imaging 2017, 36:1550-1560.

15. Bera K, Schalper KA, Rimm DL, Velcheti V, Madabhushi A: Artificial intelligence in digital pathology - new tools for diagnosis and precision oncology. Nat Rev Clin Oncol 2019, 16:703-715.

16. Zaha DC: Significance of immunohistochemistry in breast cancer. World J Clin Oncol 2014, 5:382-392.

17. Hoadley KA, Yau C, Hinoue T, Wolf DM, Lazar AJ, Drill E, Shen R, Taylor AM, Cherniack AD, Thorsson V, et al: Cell-of-Origin Patterns Dominate the Molecular Classification of 10,000 Tumors from 33 Types of Cancer. Cell 2018, 173:291-304.e296.

18. Li Y, Kang K, Krahn JM, Croutwater N, Lee K, Umbach DM, Li L: A comprehensive genomic pan-cancer classification using The Cancer Genome Atlas gene expression data. BMC Genomics 2017, 18:508.

19. Guinney J, Dienstmann R, Wang X, de Reyniès A, Schlicker A, Soneson C, Marisa L, Roepman P, Nyamundanda G, Angelino P, et al: The consensus molecular subtypes of colorectal cancer. Nat Med 2015, 21:1350-1356.

20. Roseweir AK, McMillan DC, Horgan PG, Edwards J: Colorectal cancer subtypes: Translation to routine clinical pathology. Cancer Treat Rev 2017, 57:1-7.

21. Ngiam KY, Khor IW: Big data and machine learning algorithms for health-care delivery. Lancet Oncol 2019, 20:e262-e273.

22. Weil SA, Brandt SA, Miller EL, Long DDE, Maltzahn C: Ceph: A Scalable, High-Performance Distributed File System. In Symposium on Operating Systems Design & Implementation. 2006

23. Boicea A, Radulescu F, Agapin LI: MongoDB vs Oracle -- Database Comparison. In 2012 Third International Conference on Emerging Intelligent Data and Web Technologies; 19-21 Sept. 2012. 2012: 330-335.

24. Divya MS, Goyal SK: ElasticSearch An advanced and quick search technique to handle voluminous data.

Compusoft International Journal of Advanced Computer Technology 2013, 2.

25. Lu M, Williamson D, Chen T, Chen R, Barbieri M, Mahmood F: Data Efficient and Weakly Supervised Computational Pathology on Whole Slide Images. arXiv e-prints; 2020. arXiv:2004.09666.

26. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, Thrun S: Dermatologist-level classification of skin cancer with deep neural networks. Nature 2017, 542:115-118.

27. Liu Y, Kohlberger T, Norouzi M, Dahl GE, Smith JL, Mohtashamian A, Olson N, Peng LH, Hipp JD, Stumpe MC: Artificial Intelligence-Based Breast Cancer Nodal Metastasis Detection: Insights Into the Black Box for Pathologists. Arch Pathol Lab Med 2019, 143:859-868.

28. Sadanandam A, Lyssiotis CA, Homicsko K, Collisson EA, Gibb WJ, Wullschleger S, Ostos LC, Lannon WA, Grotzinger C, Del Rio M, et al: A colorectal cancer classification system that associates cellular phenotype and responses to therapy. Nat Med 2013, 19:619-625.

29. Schlicker A, Beran G, Chresta CM, McWalter G, Pritchard A, Weston S, Runswick S, Davenport S, Heathcote K, Castro DA, et al: Subtypes of primary colorectal tumors correlate with response to targeted treatment in colorectal cell lines. BMC Med Genomics 2012, 5:66.

30. Roepman P, Schlicker A, Tabernero J, Majewski I, Tian S, Moreno V, Snel MH, Chresta CM, Rosenberg R, Nitsche U, et al: Colorectal cancer intrinsic subtypes predict chemotherapy benefit, deficient mismatch repair and epithelial-to-mesenchymal transition. Int J Cancer 2014, 134:552-562.

31. Campanella G, Werneck Krauss Silva V, Fuchs TJ: Terabyte-scale Deep Multiple Instance Learning for Classification and Localization in Pathology. arXiv e-prints; 2018. arXiv:1805.06983.

32. Frei AL, Merki S, Henke MJ, Wey N, Moch H, Mertz KD, Koelzer VH: [Future Medicine: Digital Pathology]. Ther Umsch 2019, 76:404-408.

33. Hosseini MS, Chan L, Tse G, Tang M, Deng J, Norouzi S, Rowsell C, Plataniotis KN, Damaskinos S:

Atlas of Digital Pathology: A Generalized Hierarchical Histological Tissue Type-Annotated Database for Deep Learning. In 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 15-20 June 2019. 2019: 11739-11748.

34. Marinelli RJ, Montgomery K, Liu CL, Shah NH, Prapong W, Nitzberg M, Zachariah ZK, Sherlock GJ, Natkunam Y, West RB, et al: The Stanford Tissue Microarray Database. Nucleic Acids Res 2008, 36:D871-877.

35. Falk T, Mai D, Bensch R, Çiçek Ö, Abdulkadir A, Marrakchi Y, Böhm A, Deubner J, Jäckel Z, Seiwald K, et al: U-Net: deep learning for cell counting, detection, and morphometry. Nat Methods 2019, 16:67-70.

Figures

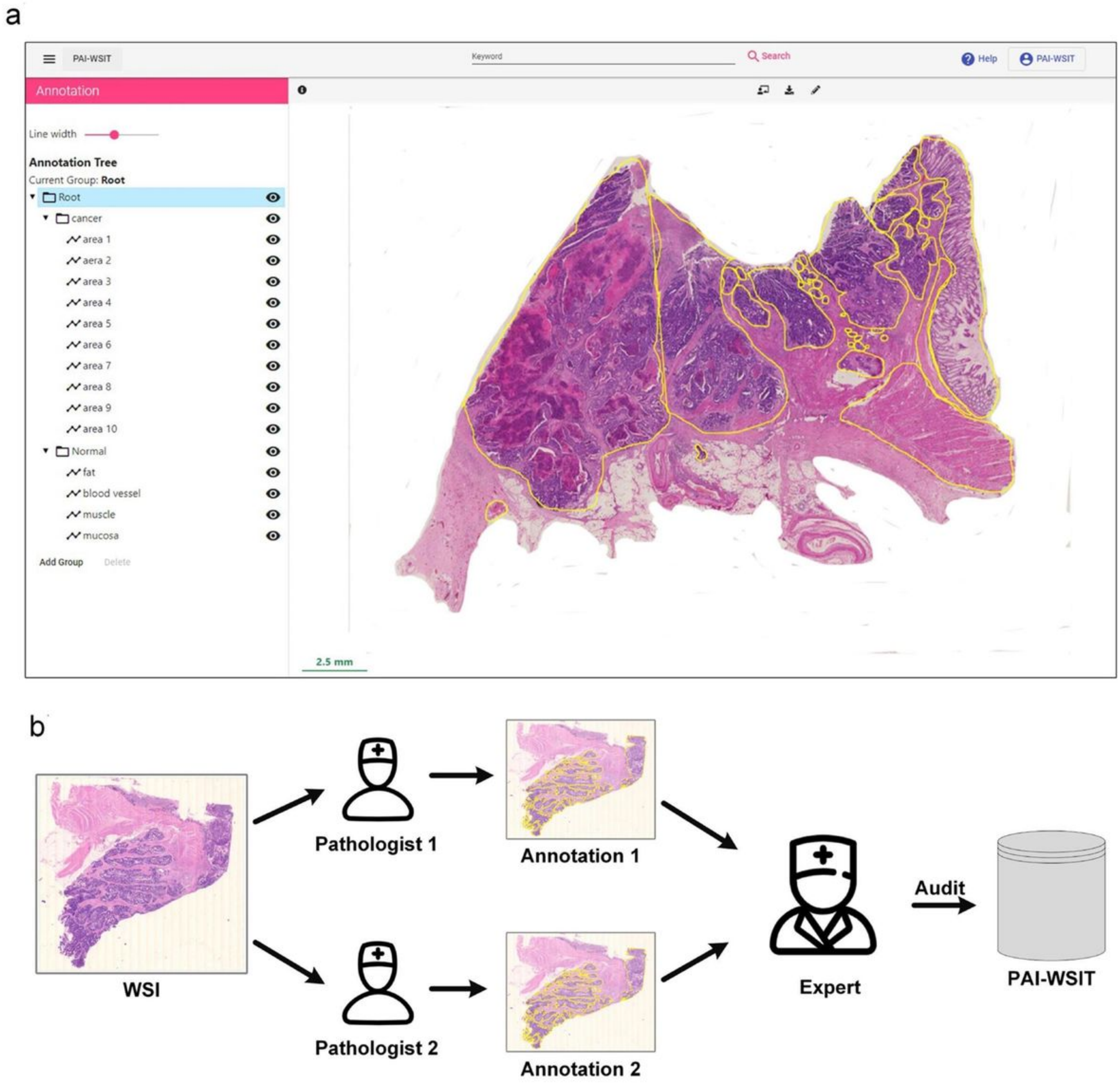

Figure 1

Annotation tool in PAI-WSIT. (a) Annotation tree: the user draws two groups of regions in a WSI, one is cancer areas and the other is normal areas. (b) Annotation protocol: each annotation in PAI-WSIT was annotated by two pathologists and audited by an experienced expert. WSI whole slide image

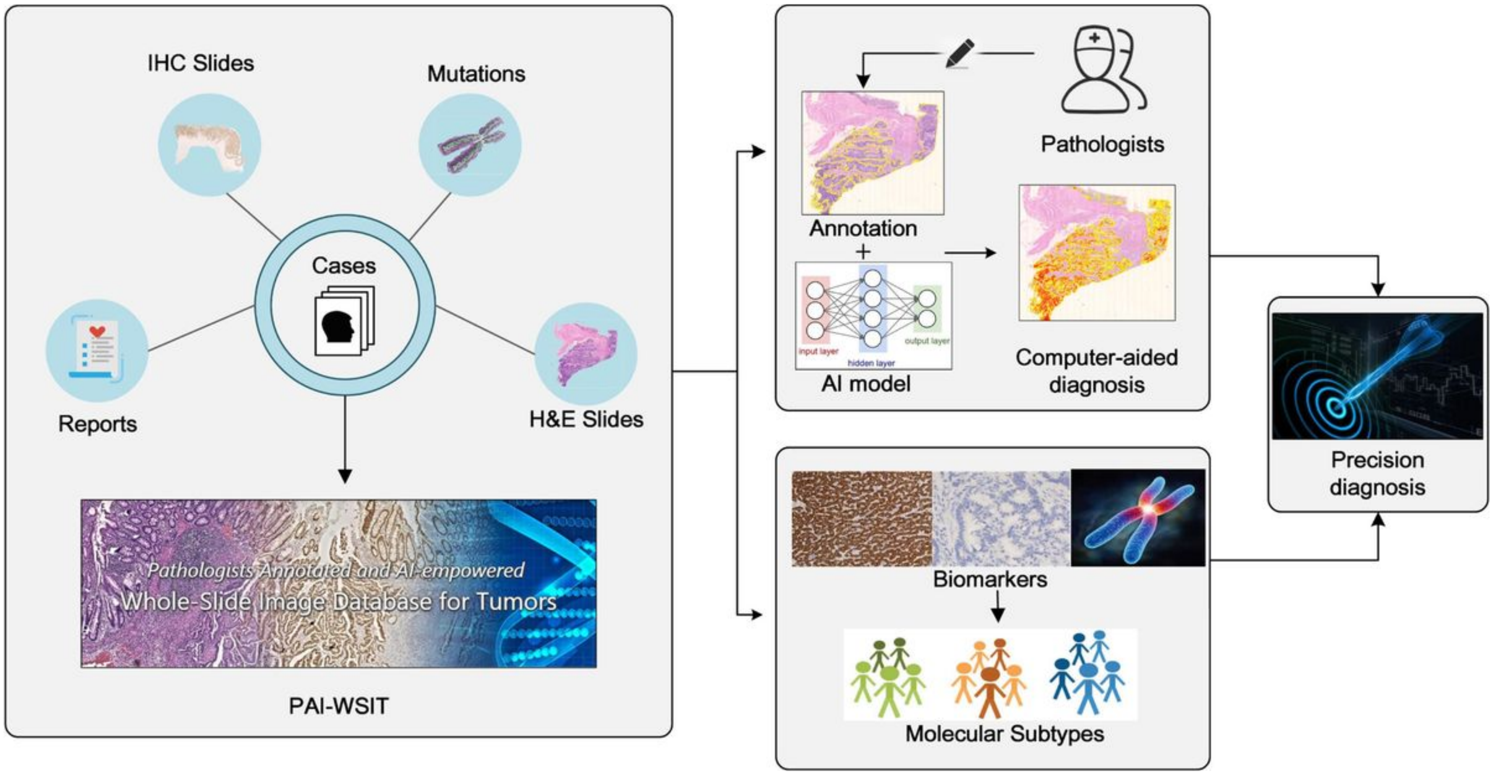

Figure 2

The overall workflow of PAI-WSIT. PAI-WSIT contains basic information, pathological reports, WSIs and target biomarker results of cancer cases. The integration of annotation tools towards WSI makes AI in digital pathology more convenient. Biomarkers in DNA-level and transcriptional-level are helpful for molecular subtype research. IHC immunohistochemistry, H&E hematoxylin-eosin, AI artificial intelligence

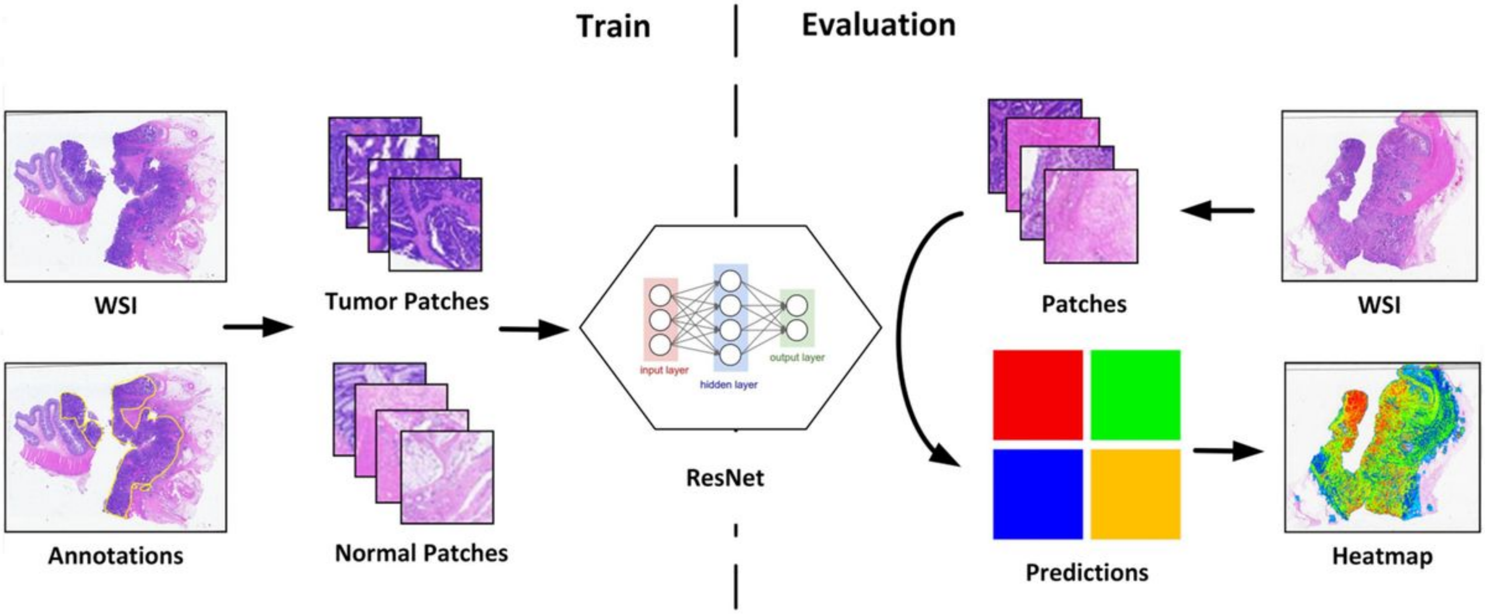

Figure 3

AI workflow. Patches were extracted by WSIs in training set, and label as positive or negative samples by annotation. After training, WSIs in validation set were predicted and visualized as heatmaps. WSI whole slide image

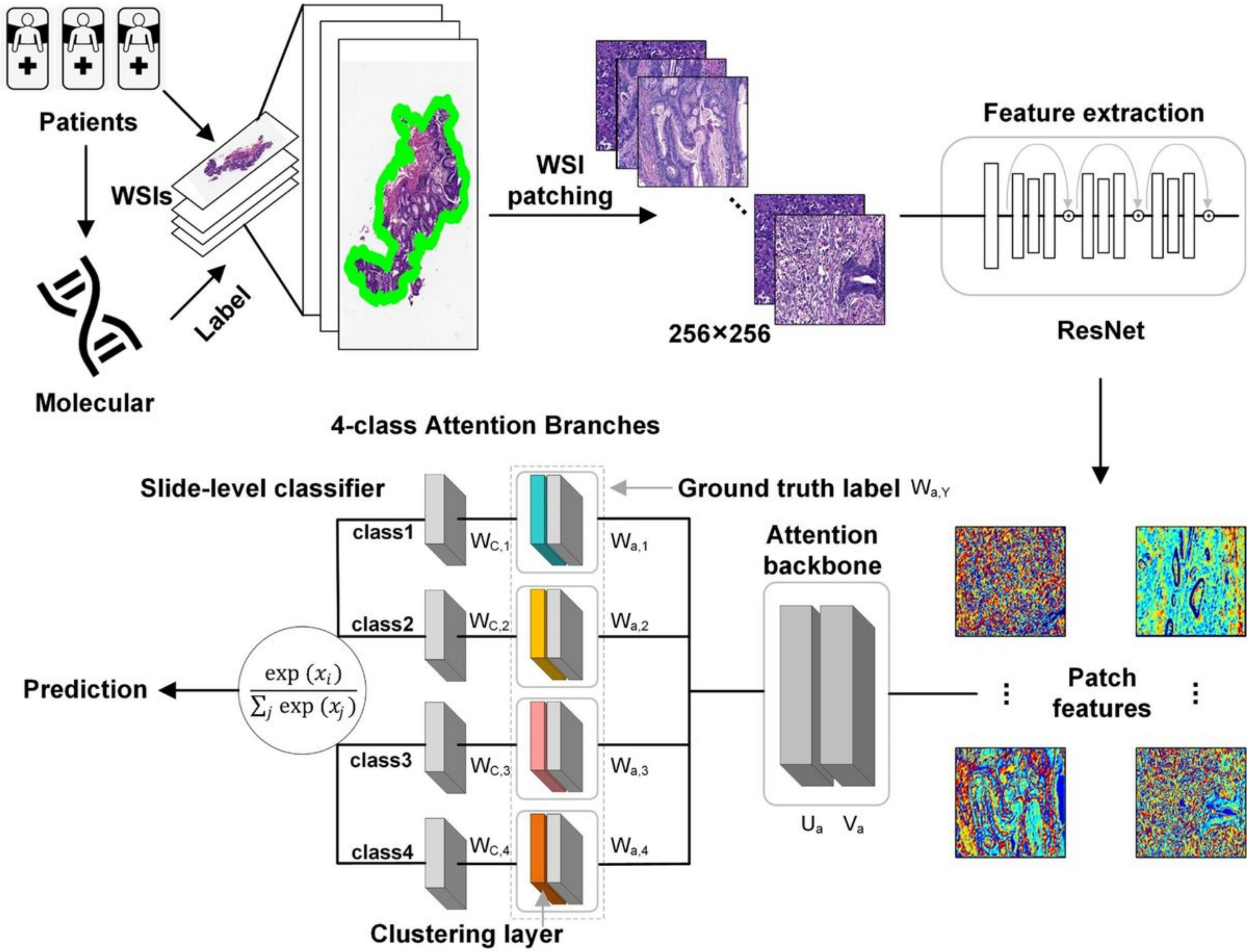

Figure 4

Pipeline of CMS classification. The WSIs of CMS were labeled by gene expression data, all patches are extracted from tissue regions of a WSI. Patches are encoded once by a pretrained ResNet into a descriptive feature representation. Attention-pooling weighs patches and summarizes patch-level features into slide-level representations to make final prediction. WSI whole slide image

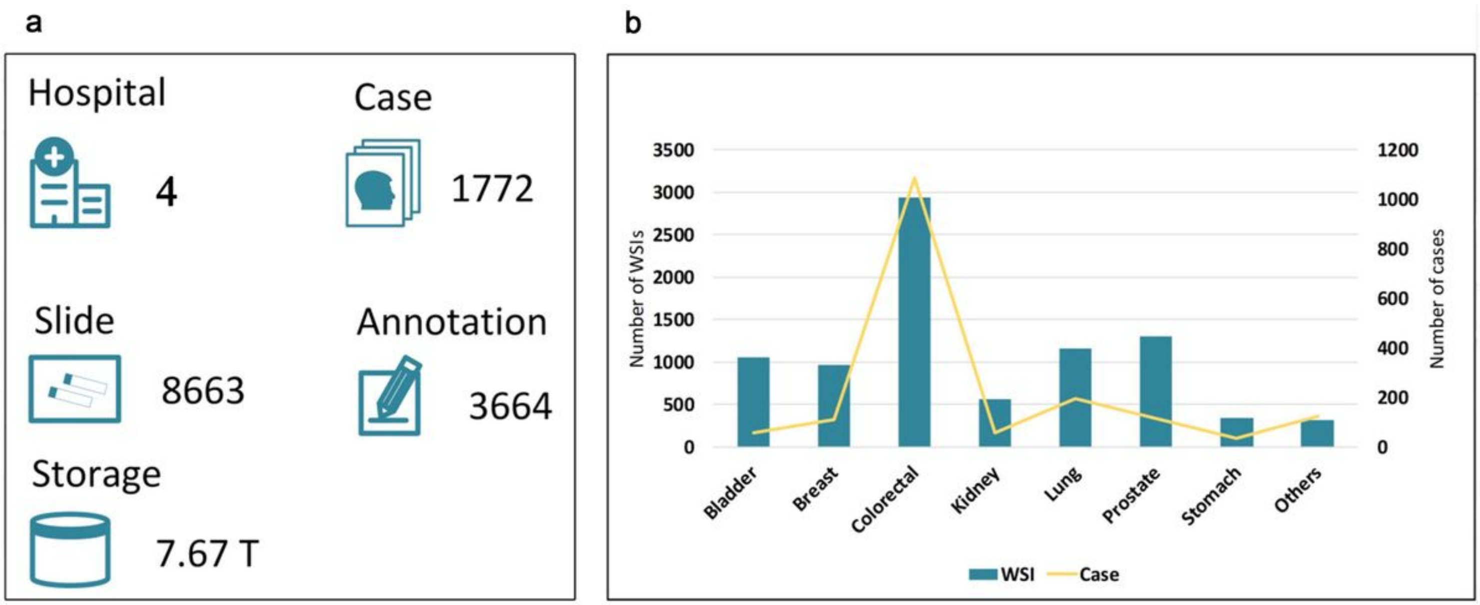

Figure 5

An overview of the data PAI-WSIT has recorded. (a) Data summary in PAI-WSIT. (b) Cases and slides by major primary site in PAI-WSIT. WSI whole slide image

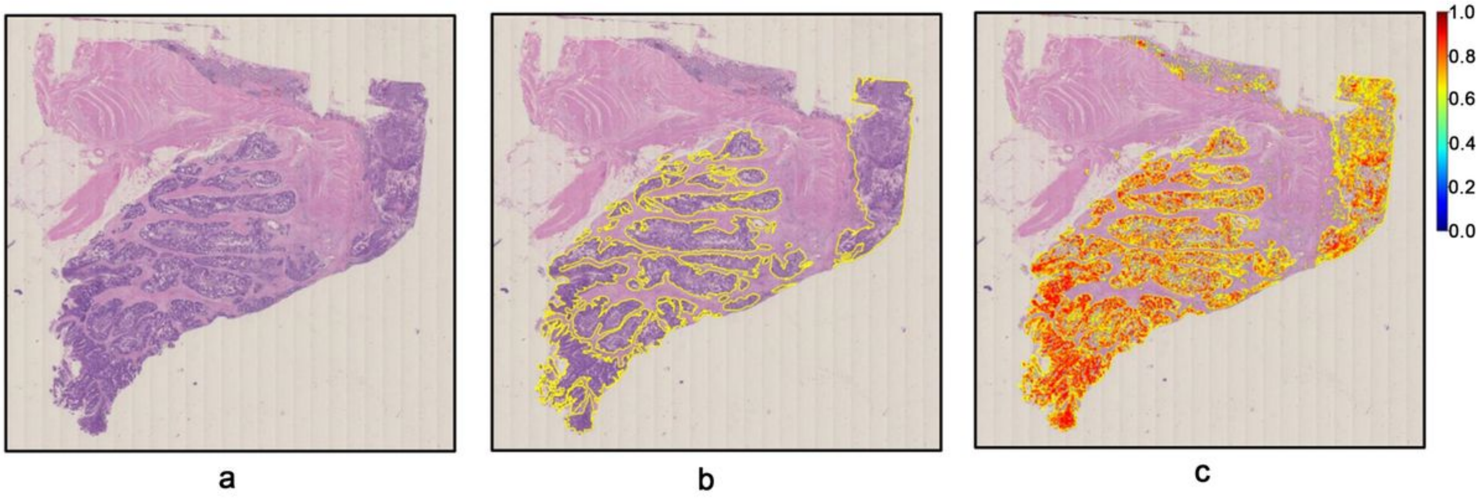

Figure 6

Prediction of location of tumor regions in ResNet. (a) Raw WSI;(b) Annotation; (c) Prediction. For each annotated slide, the whole slide attention heatmap is generated by their tumor probability. WSI whole slide image

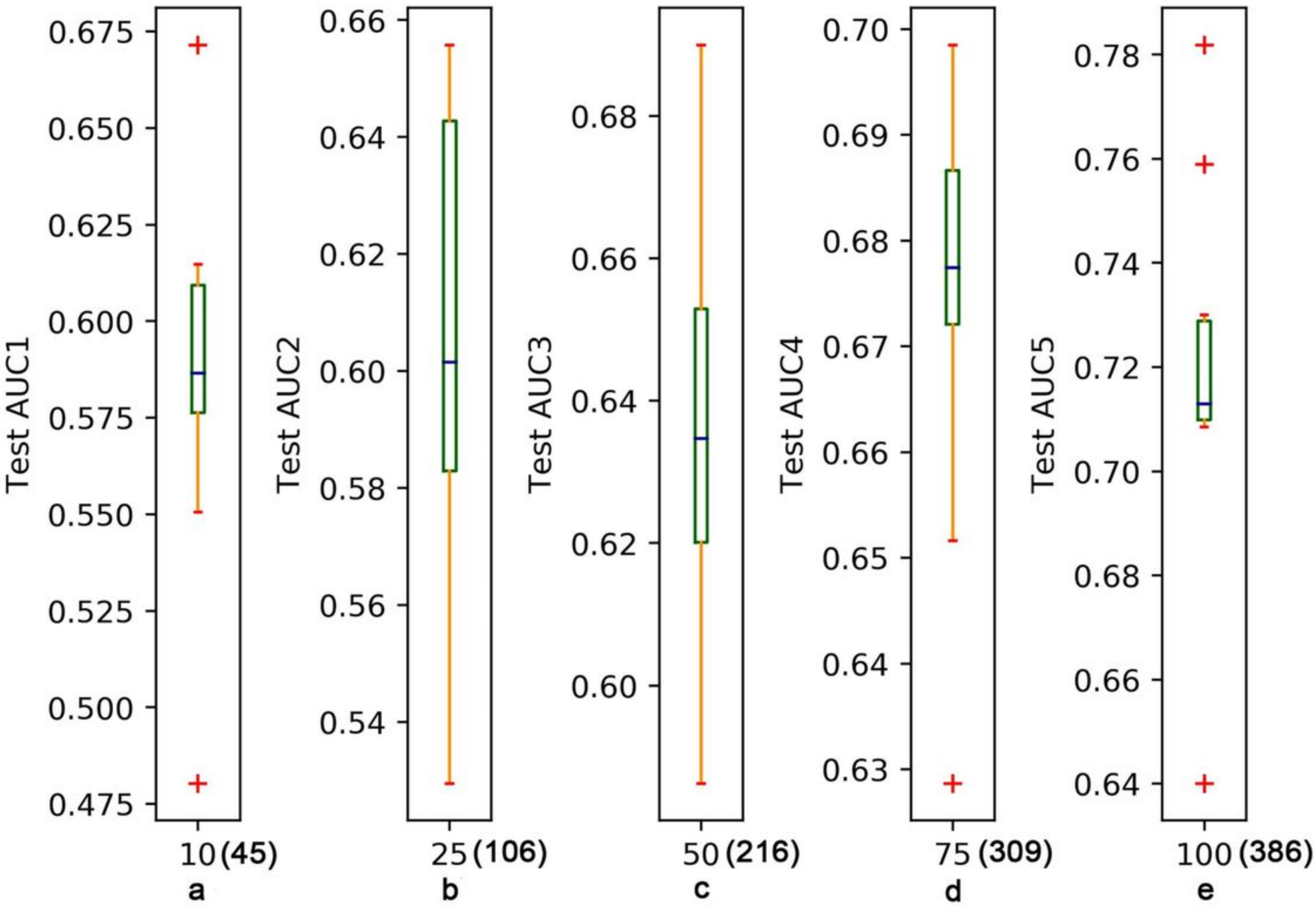

Figure 7

AUC of test set. (a-e) were % of Training set used (Number of WSIs). AUC area under the curve, WSI whole slide image

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Additionalfile1.docx
- Additionalfile2.doc