Accurate recognition of colorectal cancer with semi-supervised deep learning on pathological images

Gang Yu
Central South University
Ting Xie
Central South University
Chao Xu
University of Oklahoma Health Sciences Center https://orcid.org/0000-0002-3821-6187
Xing-Hua Shi
Temple University
Chong Wu
Florida State University https://orcid.org/0000-0002-8400-1785
Run-Qi Meng
Central South University
Xiang-He Meng
Tulane Center for Bioinformatics and Genomics, School of Public Health and Tropical Medicine, Tulane University https://orcid.org/0000-0001-8731-2899
Kuan-Song Wang (375527162@qq.com)
Central South University
Hong-Mei Xiao (hmxiao@csu.edu.cn)
Central South University
Hong-Wen Deng (hdeng2@tulane.edu)
Tulane University https://orcid.org/0000-0002-0387-8818

Research Article

Keywords: colorectal cancer, artificial intelligence, semi-supervised learning, pathological diagnosis

DOI: https://doi.org/10.21203/rs.3.rs-75912/v1

Abstract

Background: The machine-assisted recognition of colorectal cancer has been mainly focused on supervised deep learning that suffer from a significant bottleneck of requiring massive labeled data. We hypothesize that semi-supervised deep learning leveraging a small number of labeled data can provide a powerful alternative strategy.

Method: We proposed a semi-supervised model based on mean teacher that provide pathological predictions at both patch-level and patient-level. We demonstrated the general utility of the model utilizing 13,111 whole slide images from 8,803 subjects gathered from 13 centers. We compared our proposed method with the prevailing supervised learning and six pathologists.

Results: with a small amount of labeled training patches (~3,150 labeled, ~40,950 unlabeled or ~6,300 labeled,~37,800 unlabeled), the semi-supervised model performed significantly better than the supervised model (AUC: 0.90 ± 0.06 vs. 0.84 ± 0.07, *P* value = 0.02 or AUC: 0.98 ± 0.01 vs 0.92 ± 0.04, *P* value = 0.0004). Moreover, we found no significant difference between the supervised model using massive ~44,100 labeled patches and the semi-supervised model (~6,300 labeled, ~37,800 unlabeled) at patch-level diagnoses (AUC: 0.98 ± 0.01 vs 0.987 ± 0.01, *P* value = 0.134) and patient-level diagnoses (average AUC: 97.40% vs. 97.96%, *P* value = 0.117) . Our model was close to human pathologists (average AUC: 97.17% vs. 96.91%).

Conclusions: We reported that semi-supervised learning can achieve excellent performance through a multi-center study. We thus suggested that semi-supervised learning has great potentials to build artificial intelligence (AI) platforms, which will dramatically reduce the cost of labeled data and greatly facilitate the development and application of AI in medical sciences.

Introduction

Colorectal cancer (CRC) is the second most common cause of cancer death in Europe and America. [1–2]. Pathological diagnosis is one of the most authoritative methods for diagnosing CRC [3–4], which requires a pathologist to visually examine digital full-scale whole slide images (WSI). The challenges stem from the complexity of WSI including large image sizes (> 10,000 × 10,000 pixels), complex shapes, textures, and histological changes in nuclear staining [4]. Furthermore, there is a shortage of pathologists worldwide in stark contrast with the rapid accumulation of WSI data, and the daily workload of pathologists is intensive which could lead to unintended misdiagnose due to fatigue [5]. Hence, it is crucial to develop diagnosing strategies that are effective yet of low cost by leveraging recent AI development.

Deep learning provides an exciting opportunity to support and accelerate pathological analysis [6,25], including lung [7–8], breast [9–10], and skin cancers [11–12]. Progress has been made in applying deep learning to CRC including classification [13], tumor cell detection [14–15, 26], and outcome prediction [16–18]. For example, we have developed a recognition system for CRC using a supervised learning, which achieved one of highest diagnosis accuracies [19]. However, our earlier method was built upon learning from 62,919 labeled patches from 842 subjects, which were carefully selected and extensively labeled by pathologists.

While supervised learning with massive labeled data can achieve high diagnostic accuracy, the reality is that we often have only a small amount of labeled data and a much larger amount of unlabeled data. Only very few studies have investigated if semi-supervised learning, a method that leverages both labeled and unlabeled data,

can be applied to achieve satisfactory accuracy in *patient level* pathology diagnosis. For example, on a small data set of 115 WSIs, a semi-supervised method can achieve high accuracy only at *the patch level* [23]. However, to our knowledge, the CRC recognition system of semi-supervised models has not been extensively validated on *patient level* dataset from multiple centers to assess the general utility of semi-supervised learning. How to translate the patch-level prediction to WSI and patient level diagnosis is not trivial, and the patient -level diagnosis is required in clinical applications of any AI system.

To fill this gap, we used 13,111 WSIs collected from 8,803 subjects from 13 centers to develop a semi-supervised model. We evaluated the model by comparing its performance with that of prevailing supervised learning and also with that of professional pathologists. The main contributions of this paper are summarized as follows:

(1) We evaluated different CRC recognition methods based on semi-supervised and supervised learning at the patch-level and patient-level respectively. This large-scale evaluation showed that accurate CRC recognition is feasible with a high degree of reliability even when the number of labeled data is limited.

(2) We found that semi-supervised model performed better than supervised model when only a small number of labeled patches(~3,150)was available (assume a large number of unlabeled patches (e.g., ~ 40,950) available, which was often the case in practice). When ~ 6,300 labeled and ~ 37,800 unlabeled patches were used for semi-supervised training, there was no significant difference between the obtained semi-supervised model and the supervised model on ~ 44,100 labeled patches. This finding holds for CRC recognition at both the patch level and patient level.

(3) We reported that semi-supervised model (~ 6,300 labeled, ~ 37,800 unlabeled training patches) can match the accuracy of pathologists. Our study thus indicated that medical AI systems can be successfully deployed based on semi-supervised learning, and thus will dramatically reduce the amount of labeled data required in practice, to greatly facilitate the development and application of AI in medical sciences.

Results

We trained and tested our method utilizing CRC datasets from multiple centers (Fig. 1, Table 1). Briefly, we divided each WSI into thousands of patches. At the patch level, we applied a semi-supervised learning strategy called the mean teacher [21], where a teacher network provided pseudo labels for unlabeled images participating in training. At the WSI and patient level, we applied a cluster-based and positive sensitivity strategy to achieve CRC diagnosis for patients as we did recently [19]. Refer to Supplementary A for details of methodology.

Semi-supervised vs supervised recognition at patch level

The 62,919 patches in Dataset-PATT (Table 2) were used for patch-level training and testing. For simplicity, we used SSL, SL to represent semi-supervised and supervised learning methods, and a numerical number to represent the proportion of labels of the total 62,919 patches which led to the five models described as follows. Model-5%-SSL and model-10%-SSL were trained on 5% (~ 3,150) and 10% (~ 6,300) labeled patches, respectively, where the remained patches (~ 40,950 and ~ 37,800) were used, but their labels were ignored. Model-5%-SL (supervised learning) and model-10%-SL were trained on the same labeled patches only with model-5%-SSL and model-10%-SSL respectively, without using the remained patches (as unlabeled). Model-70%-SL used ~ 44,100 labeled training patches (70% of 62,919). Refer to Table 3 for details.

The AUC and 75% confidence interval were shown in Table 4 and Figure 2. With a very small amount (~ 3,150) of labeled training patches, model-5%-SSL (with ~40,950 unlabeled patches) was superior to model-5%-SL (AUC (Area Under the Curve): 0.90 ± 0.06 vs. 0.84 ± 0.07, *P* value = 0.02). With the availability of ~6,300 labeled and ~37,800 unlabeled patches, the model-10%-SSL was also significantly better than model-10%-SL (AUC: 0.98 ± 0.01 vs. 0.92 ± 0.04, *P* value = 0.0004).

The performance of model-10%-SSL (with ~6,300 labeled and ~37,800 unlabeled training patches) had no significant difference with that of the model-70%-SL (with ~44,100 labeled training patches) (AUC: 0.98 ± 0.01 vs. 0.987 ± 0.01, *P* value = 0.134). Visual inspection (Supplementary Figure 2) confirmed that that model-10%-SL failed to identify the pixels of cancer in the patches, while the pixels of cancer identified by model-10%-SSL and model-70%-SL were highly matched.

Patient-level CRC recognition

To test whether the above conclusion at patch-level still holds at patient level, we further conducted comparisons using Dataset-PT. As illustrated in Fig. 3 and Supplementary Table 2, we found that model-10%-SSL had a significant improvement over model-10%-SL (Average AUC: 97.40% vs. 81.88%, *P* value = 0.0022) on patient-level prediction in the multi-centers scenario. The average AUC of model-10%-SSL was slightly lower than, but comparable to, that of model-70%-SL (Average AUC: 97.40% vs. 97.96%, *P* value = 0.117). Among the 7 datasets (XH-dataset-PT, XH-dataset-HAC, PCH, TXH, FUS, SWH, TCGA, 11,290 WSIs), the AUC difference of model-10%-SSL and model-70%-SL was smaller than 1.6%. In particular, on the largest dataset, XH-dataset-PT (10,003 WSIs), the AUCs of model-10%-SSL and model-70%-SL were close with 98.41% vs. 99.16%. On the HPH, SYU, CGH and AMU (501 WSIs), the AUCs of model-10%-SSL were even higher than that of model-70%-SL.

In the data from GPH, and ACL (392 WSIs), the performance of model-10%-SSL was lower than that of model-70%-SL (AUC DIFF > 2.22%). It is worth noting that model-10%-SSL generally achieved good sensitivity, which proved practically useful for the diagnosis of CRC. Visual inspection in Supplementary Fig. 3 showed the cancer patches identified by model-10%-SSL and model-70%-SL were the true cancer locations on WSIs.

Human-AI competition

To evaluate the model performances for practical clinical applications, we recruited six pathologists with 1–18 years of independent experience (Supplementary Table 3). They independently reviewed 1,634 WSIs from 10 data centers (Dataset-HAC, Fig. 4).

We ranked pathologists, model-10%-SSL and model-70%-SL. The average AUC of model-10%-SSL was 97.17%, ranked at the 5th, which was close to the average AUC of pathologists (96.91%). The sensitivity of model-10%-SSL was 97.68%, showing an excellent detection ability of cancer (Supplementary Table 5).

Discussion

Accurately diagnosing CRC requires years of training, leading to a global shortage of pathologists [2]. Almost all existing computer-assisted diagnosis models currently relies on massive labeled data with supervised learning approach, but manual labeling is usually time-consuming and costly. This leads to an increasing interest in building an accurate diagnosis system with far less labeled data plus the ever-increasing unlabeled data.

In this study, we developed a semi-supervised learning method for CRC diagnosis, and evaluated its performance using an extensive collection of WSIs across 13 medical centers. On this large data set, we conducted a range of comparison of CRC recognition performance among semi-supervised learning, supervised learning and six human pathologists, at both patch level and patient level.

We demonstrated that semi-supervised learning outperformed supervised learning at patch-level recognition when only a small amount of labeled and large amounts of unlabeled data were available. In our previous study [19], we used 62,919 labeled patches from 842 WSIs, which achieved accurate patch-level recognition. When semi-supervised learning was used as demonstrated in this study, only about a tenth (6,300) of those many labeled patches plus 37,800 unlabeled patches were used to achieve similar AUC to [19] (i.e. model-70%-SL).

We also conducted extensive testing of three models for patient level prediction on 12 centers (Dataset-PT). Just like the patch level, at the patient level, the semi-supervised model outperformed the supervised model when a small number of labeled patches was available, and close to the supervised model when using a large number of labeled patches. The AUC of model-10%-SL was 96.44%, perhaps because both the testing data and training data were from XH-Dataset-PT.

However, using the data from 12 centers, the average AUC of model-10%-SL was dramatically reduced to 81.88% from 96.44% in XH-Dataset-PT. This result showed that when training data and testing data were not the same source, the generalization performance of model-10%-SL was significantly reduced. The cancerous prediction of model-10%-SL cannot be extended to other centers. Moreover, many cancerous patches predicted by model-10%-SL was deviated from true cancer locations in a WSI (Supplementary Fig. 3).

When a large number of unlabeled patches was added for model-10%-SSL, the generalization performance across centers can be maintained, where there was no significant difference when comparing with model-70%-SL using massive labeled patches. These results showed that when labeled patches were seriously insufficient, using unlabeled data can greatly improve the generalization ability across different data sets. The patient-level results indicated that with semi-supervised learning, we may not need as much labeled data as in supervised learning. Since it is well known that unlabeled medical data are relatively easy to obtain, it is of great importance and with an urgent need to develop semi-supervised learning methods.

We compared the diagnosis of six pathologists from our semi-supervised model. We found that our semi-supervised model reached an average AUC of pathologists, which was approximately equivalent to a pathologist with five years of clinical experience. The Human-AI competition in this regard thus showed that it was feasible to build an expert-level method for clinical practice based on semi-supervised learning approach.

In practice, the exact amount of the data that needs to be labeled is generally unknown. Nonetheless, as shown in our experiments, it is an alternative low-cost approach to conduct semi-supervised training with a small amount of labeled data plus a large amount of unlabeled data. Hence, it is an effective strategy to wisely utilize all data so that a small amount of data is first labeled to build a baseline model based on a semi-supervised learning. If the results are not satisfactory for this baseline model, the amount of labeled data should be increased. This strategy is feasible since as expected, semi-supervised learning requires a much smaller number of labeled data to achieve the same performance compared with a supervised learning method.

Although studies have shown that semi-supervised learning achieved nice results in tasks like natural image processing [22], semi-supervised learning has not been widely evaluated for analyzing pathological images. It is unclear whether existing semi-supervised methods can overcome the limitation of insufficient labeled pathological images. Our work confirmed that unlabeled data could improve CRC recognition. As demonstrated in our study, semi-supervised learning has excellent potentials to overcome the bottleneck of insufficient labeled data as in many medical domains.

Conclusion

Currently, patient-level computer-assisted CRC diagnosis is solely based on supervised learning, which requires a large number of labeled data to achieve good performance. In this study, we applied a semi-supervised method and extensively evaluated its performance on multi-center datasets. We demonstrated that semi-supervised learning with a small number of labeled data achieved comparable prediction accuracy as that of supervised learning with massive labeled data and that of experienced pathologists. This study thus supported potential applications of semi-supervised learning to develop medical AI systems.

Declarations

Acknowledgement

K.S.W was supported by the National Natural Science Foundation of China (#81673491) , Natural Science Foundation of Hunan Province (#2015JJ2150). H.M.X was supported by the National Key Research and Development Plan of China (2017YFC1001103, 2016YFC1201805), National Natural Science Foundation of China (#81471453), and Jiangwang Educational Endowment. H.W.D. were supported by the National Institutes of Health (R01AR059781, P20GM109036, R01MH107354, R01MH104680, R01GM109068, R01AR069055, U19AG055373, R01DK115679), the Edward G. Schlieder Endowment and the Drs. W.C.Tsai and P.T.Kung Professorship in Biostatistics from Tulane University.

References

1. Arnold M, Sierra MS, Laversanne M, Soerjomataram I, Jemal A, Bray F. Global patterns and trends in colorectal cancer incidence and mortality. Gut 2017; 66:683–691.

2. Metter DM, Colgan TJ, Leung ST, Timmons CF, Park JY. Trends in the US and Canadian Pathologist Workforces From 2007 to 2017. JAMA Netw Open 2019;2: e194337.

3. Ivan Damjanov. Robbins Review of Pathology[J]. Modern Pathology, 2000, 13(9):1028–1028.

4. Group C C C W. Chinese Society of Clinical Oncology (CSCO) diagnosis and treatment guidelines for colorectal cancer 2018 (English version) [J]. Chinese Journal of Cancer Research, 2019, 31(1): 99–116.

5. Sayed S, Lukande R, Fleming KA. Providing Pathology Support in Low-Income Countries. J Glob Oncol 2015; 1:3–6.

6. Komura D, Ishikawa S. Machine Learning Methods for Histopathological Image Analysis. Comput Struct Biotechnol J. 2018; 16:34–42.

7. Coudray N, Ocampo PS, Sakellaropoulos T, et al. Classification and mutation prediction from non-small cell lung cancer histopathology images using deep learning. Nat Med. 2018;24(10):1559–1567.

8. Hua KL, Hsu CH, Hidayati SC, Cheng WH, Chen YJ. Computer-aided classification of lung nodules on computed tomography images via deep learning technique. Onco Targets Ther. 2015 Aug 4;8:2015-22.

9. Veta M, van Diest PJ, Willems SM, et al. (2015). Assessment of algorithms for mitosis detection in breast cancer histopathology images. Med Image Anal 20, 237–248.

10. Ehteshami Bejnordi B, Veta M, Johannes van Diest P, et al. Diagnostic Assessment of Deep Learning Algorithms for Detection of Lymph Node Metastases in Women with Breast Cancer. JAMA. 2017;318(22):2199–2210.

11. Zhang N, Cai Y X, Wang Y Y, et al. Skin Cancer Diagnosis Based on Optimized Convolutional Neural Network[J]. Artificial Intelligence in Medicine, 2019, 102:101756.

12. Andre Esteva, Brett Kuprel, Roberto A. Novoa, et al, Dermatologist-level classification of skin cancer with deep neural networks, Nature,2017, 542(2): 115–126.

13. Haj-Hassan, H., Chaddad, A., Harkouss, Y., Desrosiers, C., Toews, M., and Tanougast, C. Classifications of Multispectral Colorectal Cancer Tissues Using Convolution Neural Network. J Pathol Inform,2017,8: 1.

14. Sirinukunwattana K, Ahmed Raza SE, Yee-Wah T, Snead DR, Cree IA, Rajpoot NM. Locality Sensitive Deep Learning for Detection and Classification of Nuclei in Routine Colon Cancer Histology Images. IEEE Trans Med Imaging 2016; 35:1196 – 206.

15. Chaddad A, Tanougast C. Texture Analysis of Abnormal Cell Images for Predicting the Continuum of Colorectal Cancer. Anal Cell Pathol (Amst) 2017; 2017:8428102.

16. Bychkov D, Linder N, Turkki R, et al. Deep learning based tissue analysis predicts outcome in colorectal cancer. Sci Rep 2018; 8:3395.

17. Kather JN, Krisam J, Charoentong P, et al. Predicting survival from colorectal cancer histology slides using deep learning: A retrospective multicenter study. PLOS Medicine 2019;16:e1002730.

18. Skrede OJ, De Raedt S, Kleppe A, et al. Deep learning for prediction of colorectal cancer outcome: a discovery and validation study. The Lancet, 2020, 395(10221), 350–360.

19. Wang Kuan-Song, Yu Gang, Xu Chao, et al, Accurate Diagnosis of Colorectal Cancer Based on Histopathology Images Using Artificial Intelligence. bioRxiv preprint: 10.1101/2020.03.15.992917.

20. ari CT, Gunduz-Demir C. Unsupervised Feature Extraction via Deep Learning for Histopathological Classification of Colon Tissue Images. IEEE Trans Med Imaging. 2019;38(5):1139–1149.

21. Antti Tarvainen, Harri Valpola, Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results, arXiv preprint arXiv:1703.01780v6

22. I Zeki Yalniz, Herv´e J´egou, Kan Chen, Manohar Paluri, and Dhruv Mahajan. Billion-scale semisupervised learning for image classification. arXiv preprint arXiv:1905.00546, 2019.

23. Shayne Shaw, Maciej Pajak, Aneta Lisowska, Sotirios A. Tsaftaris, Alison Q. ONel, Teacher-student chain for efficient semi-supervised histology image classification, arXiv preprint arXiv:2003.08797v2, 2020.

24. Szegedy C, Wei L, Yangqing J, et al. Going Deeper with Convolutions. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, 7–12 June 2015, 1–9.

25. Campanella G, Hanna MG, Geneslaw L, et al. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. Nature Medicine 2019.

26. Wei JW, Suriawinata A A, Vaickus LJ, et al. (2019). Deep neural networks for automated classification of colorectal polyps on histopathology slides: a multi-institutional evaluation, arXiv preprint arXiv: 1909.12959v2, 2019.

Tables

Table 1. Datasets used from multi-center data sources

Data source	Dataset Usage	Sample preparation	Examination type	Population	CRC		Non-CRC		Total	
			Radical surgery / Colonoscopy		subjects	slides	subjects	slides	subjects	slides
Xiangya Hospital (XH)	PATT	FFPE	100% / 0%	Changsha, China	614	614	228	228	842	842
NCT-UMM (NCT-CRC-HE-100K)	PAT	FFPE	NA	Germany	NA	NA	NA	NA	NA	86
Xiangya Hospital (XH-dataset-PAT)	PT	FFPE	80% / 20%	Changsha, China	3,990	7,871	1,849	2,132	5,839	10,003
Xiangya Hospital (XH-dataset-HAC)	HAC	FFPE	89% / 11%	Changsha, China	98	99	97	114	195	213
Pingkuang Collaborative Hospital (PCH)	PT & HAC	FFPE	60% / 40%	Jiangxi, China	50	50	46	46	96	96
The Third Xiangya Hospital of CSU (TXH)	PT & HAC	FFPE	61% / 39%	Changsha, China	48	70	48	65	96	135
Hunan Provincial People's Hospital (HPH)	PT& HAC	FFPE	61% / 39%	Changsha, China	49	50	49	49	98	99
Adicon clinical laboratory (ACL)	PT & HAC	FFPE	22% / 78%	Changsha, China	100	100	107	107	207	207
Fudan University Shanghai Cancer Center (FUS)	PT & HAC	FFPE	97% / 3%	Shanghai, China	100	100	98	98	198	198
Guangdong Provincial People's Hospital (GPH)	PT & HAC	FFPE	77% / 23%	Guangzhou, China	100	100	85	85	185	185
Southwest Hospital (SWH)	PT & HAC	FFPE	93% / 7%	Chongqing, China	99	99	100	100	199	199
The First Affiliated Hospital Air Force Medical University (AMU)	PT & HAC	FFPE	95% / 5%	Xi'an, China	101	101	104	104	205	205
Sun Yat-Sen University Cancer Center (SYU)	PT & HAC	FFPE	100% / 0%	Guangzhou, China	91	91	6	6	97	97
Chinese PLA General Hospital (CGH)	PT	FFPE	NA	Beijing, China	0	0	100	100	100	100
The Cancer Genome Atlas (TCGA-FFPE)	PT	FFPE	100% / 0%	U.S.	441	441	5	5	446	446
Total					5,881	9,786	2,922	3,239	8,803	13,111

PATT: patch-level training and test. PAT: independent patch-level test. PT: patient-level test. HAC: human-AI competition. XH-dataset-PAT: XH data in dataset-PAT. XH-dataset-HAC: XH data in dataset-HAC.

NCT-UMM: https://zenodo.org/record/1214456#.XV2cJeg3lhF. The TCGA data were downloaded at https://portal.gdc.cancer.gov/.

Table 2. Dataset-PATT and Dataset-PAT

Dataset	Cancer			Non-cancer			Total		
	subjects	slides	patches	subjects	slides	patches	subjects	slides	patches
Dataset-PATT	614	614	30056	228	228	32863	842	842	62919
Dataset-PAT	NA	NA	14,317	NA	NA	85,683	NA	86	100,000
Total	>614	>614	44,373	>228	>228	118,546	>842	928	162,919

Table 3. Training and testing sets for patch-level models

Model	Dataset-PATT (training)			Dataset-PATT (test)		Dataset-PAT	
	Cancer	Non-cancer	unused label	cancer	Non-cancer	cancer	Non-cancer
Model-5%-SSL	5%	5%[a]	65%[d]	30%	30%	14317	85683
Model-10%-SSL	10%	10%[b]	60%[e]	30%	30%	14317	85683
Model-5%-SL	5%	5%[a]	-	30%	30%	14317	85683
Model-10%-SL	10%	10%[b]	-	30%	30%	14317	85683
Model-70%-SL	70%	70%[c]	-	30%	30%	14317	85683

a-e: About 3,150, 6,300, 44,100, 40,950, 37,800 patches from 5%-60% patients, and there are no too many patches extracted from any patient.

Table 4. AUC and 75% Confidence interval of two test sets

Model	Dataset-PATT (test)	Dataset-PAT	Both sets	P value[a]
Model-5%-SSL	0.90 ± 0.08	0.90 ± 0.02	0.90 ± 0.06	0.02
Model-5%-SL	0.79 ± 0.02	0.89 ± 0.04	0.84 ± 0.07	
Model-10%-SSL	0.99 ± 0.01	0.97 ± 0.01	0.98 ± 0.01	0.0004
Model-10%-SL	0.94 ± 0.04	0.91 ± 0.03	0.92 ± 0.04	
Model-10%-SSL	0.99 ± 0.01	0.97 ± 0.01	0.98 ± 0.01	0.134
Model-70%-SL	0.994 ± 0.01	0.98 ± 0.01	0.987 ± 0.01	

a: Wilcoxon signed rank test

Figures

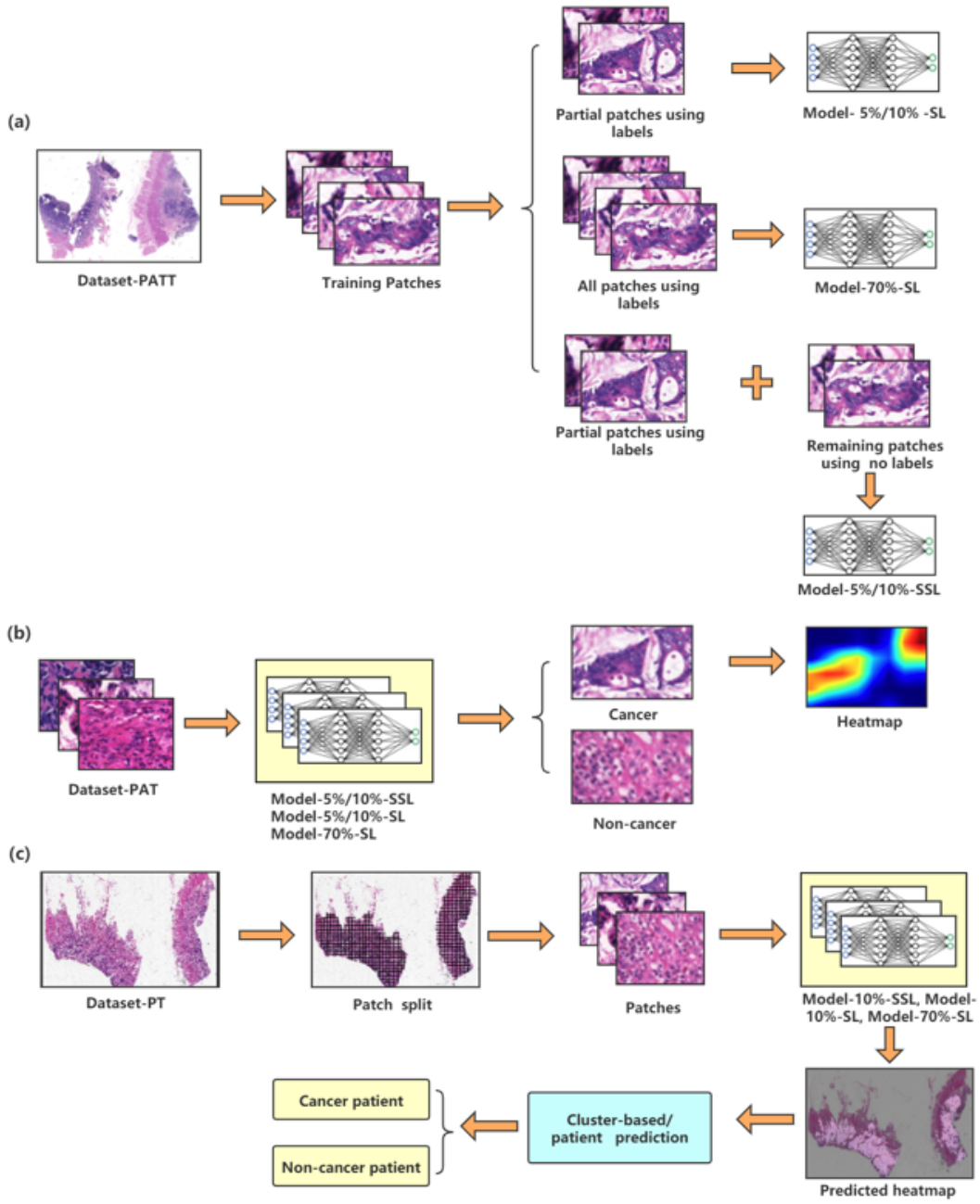
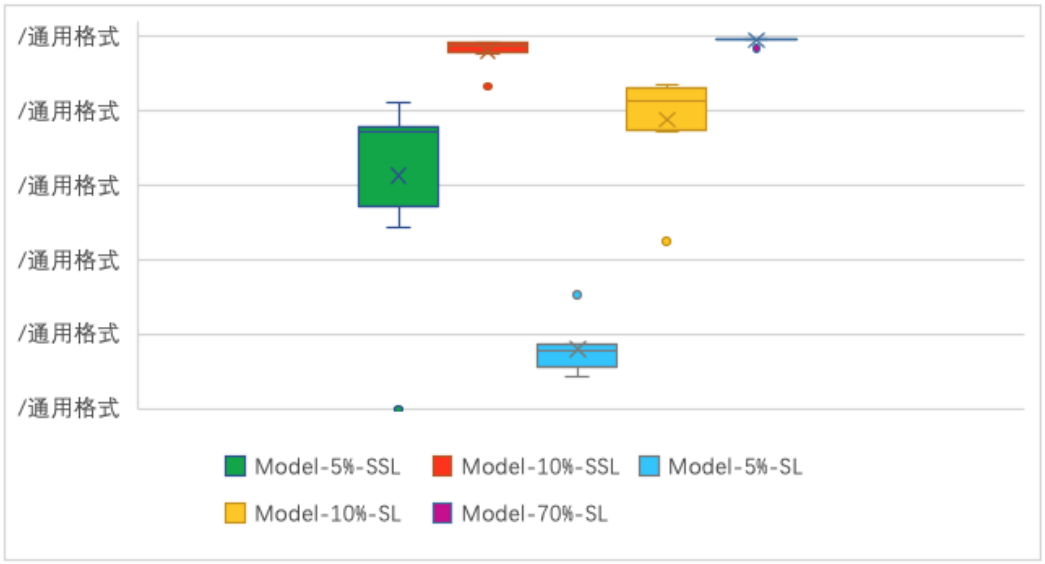
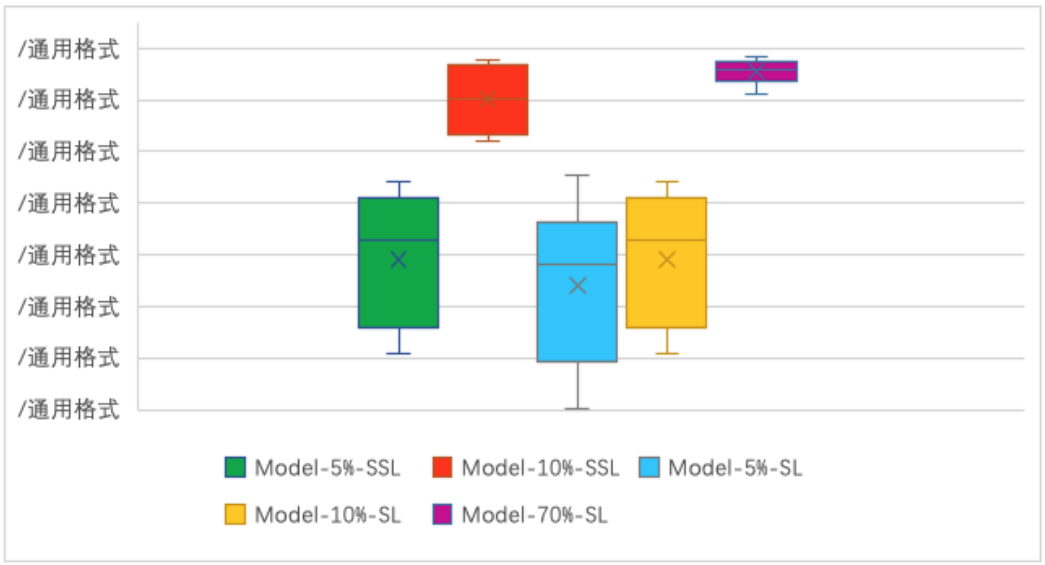


Figure 1

The flow chart of the study. (a) Semi-supervised and supervised training are performed on Dataset-PATT. (b) The patch-level test of five models on Dataset-PAT. (c) The patient-level test used Dataset-PT. The heatmap shows the cancer locations in WSI.



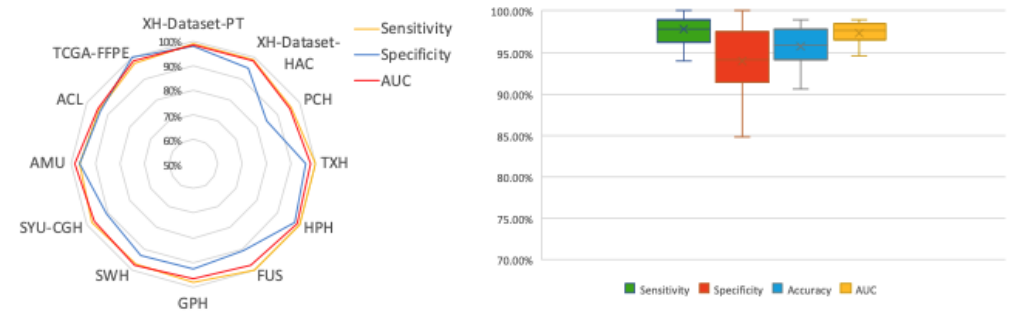
(a) Dataset-PATT (test)



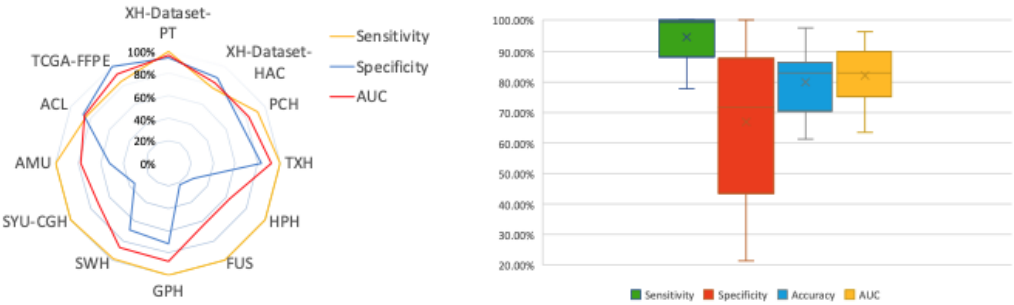
(b) Dataset-PAT

Figure 2

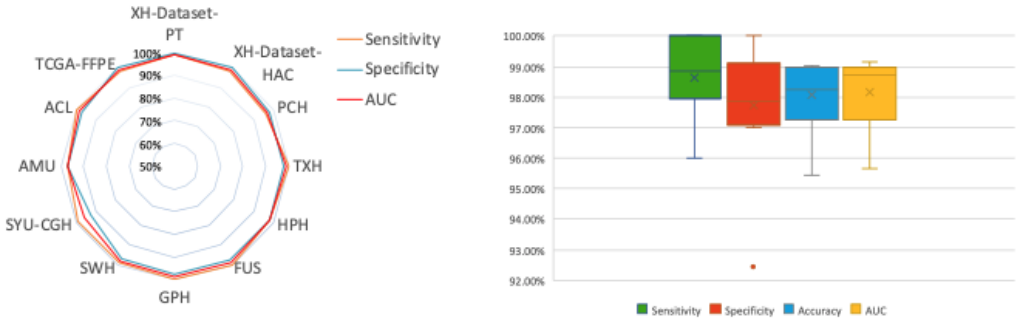
The AUC distribution of five models at patch level on two datasets.



(a) Model-10%-SSL



(b) Model-10%-SL



(c) Model-70%-SL

Figure 3

Patient-level comparison on twelve independent datasets. Left: Radar maps illustrating the sensitivity, specificity, and AUC. Right: Boxplots showing the distribution of sensitivity, specificity, accuracy, and AUC in these datasets.

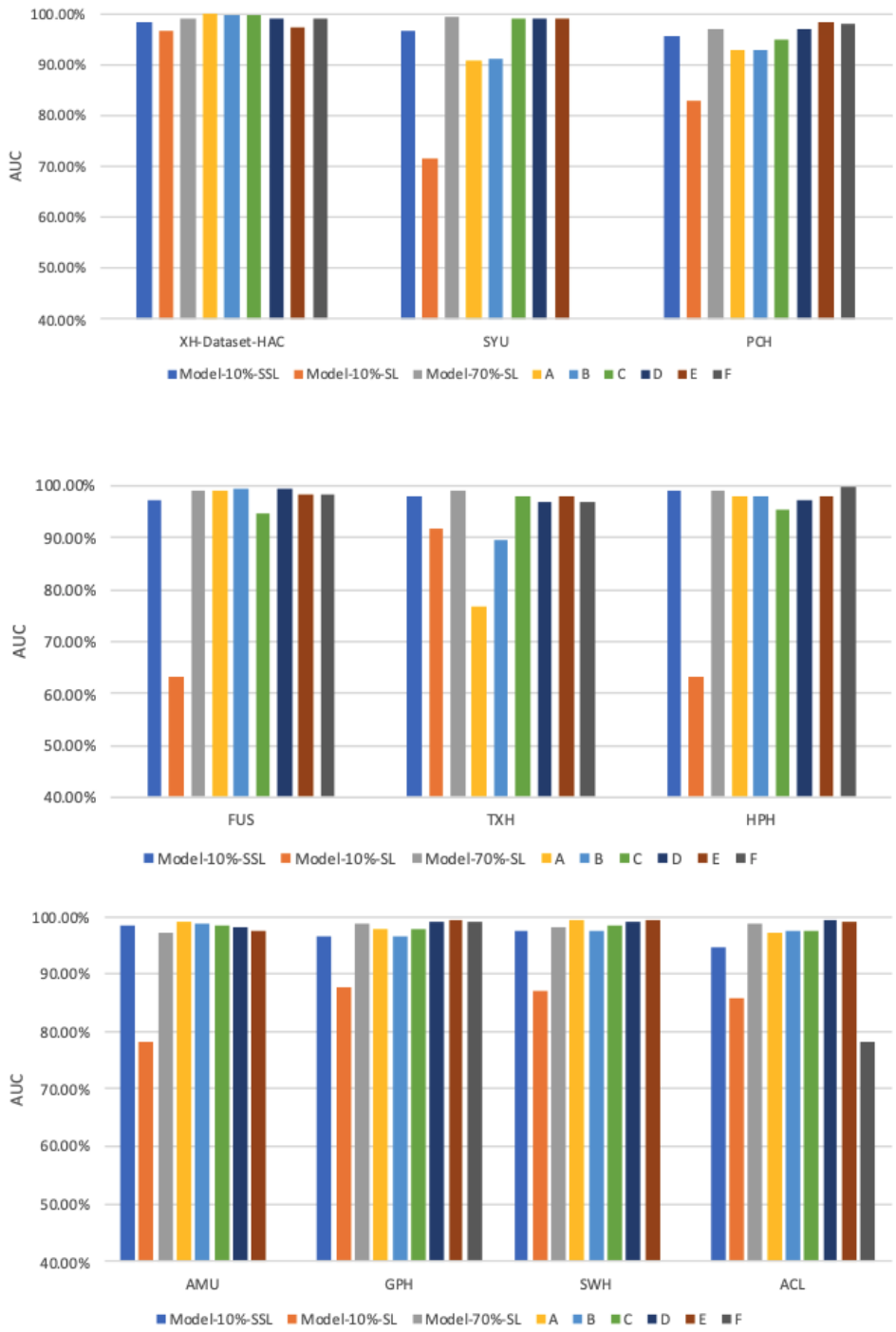


Figure 4

AUC comparison of in the Human-AI contest using Dataset-HAC. Colored lines indicate the AUCs achieved by three models and six pathologists (A-F).

Supplementary Files

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

- .Accuraterecognitionofcolorectalcancerwithsemisupervisedeeplearningonpathologicalimages.docx