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

Gang Yu[1], Ting Xie[1], Chao Xu[2], Xing-Hua Shi[3], Chong Wu[4], Run-Qi Meng[5], Xiang-He Meng[6], Kuan-Song Wang[7, #], Hong-Mei Xiao[6, #], Hong-Wen Deng[6,8, #]

1. Department of Biomedical Engineering, School of Basic Medical Science, Central South University, Changsha, Hunan, 410013, China.
2. Department of Biostatistics and Epidemiology, University of Oklahoma Health Science Center, Oklahoma City, OK, 73104, USA.
3. Department of Computer & Information Sciences, College of Science and Technology, Temple University, Philadelphia, PA, 19122, USA.
4. Department of Statistics, Florida State University, Tallahassee, FL, 32306, USA.
5. Electronic Information Science and Technology, School of Physics and Electronics, Central South University, Changsha, Hunan, 410083, China.
6. School of Basic Medical Science, Central South University, Changsha, Hunan, 410013, China.
7. Department of Pathology, Xiangya Hospital, Central South University, Changsha, Hunan, 410078, China.
8. Tulane Center of Biomedical Informatics and Genomics, Deming Department of Medicine, Tulane University School of Medicine, New Orleans, LA, 70112, USA.

#Co-corresponding authors:
Kuan-Song Wang, Ph.D.
Associate Professor
Department of Pathology
Xiangya Hospital
Central South University
Changsha, Hunan, 410078, P.R. China
Email: 375527162@qq.com

Hong-Mei Xiao, M.D., Ph.D.
Professor, School of Basic Medical Sciences
Central South University
Changsha, Hunan 410000, P.R. China
Email: hmxiao@csu.edu.cn

Hong-Wen Deng, Ph.D.
Professor
Director, Tulane Center for Biomedical Informatics and Genomics
Deming Department of Medicine
Tulane University School of Medicine
1440 Canal Street, Suite 2001, New Orleans, LA 70112, USA
Tel: 504-988-1310; Fax: 504-988-1706
Email: hdeng2@tulane.edu

Abstract

Purposes: The machine-assisted recognition of colorectal cancer using pathological images has been mainly focused on supervised learning approaches that suffer from a significant bottleneck of requiring a large number of labeled training images. The process of generating high quality image labels is time-consuming, labor-intensive, and thus lags behind the quick accumulation of pathological images. We hypothesize that semi-supervised deep learning, a method that leverages a small number of labeled images together with a large quantity of unlabeled images, can provide a powerful alternative strategy for colorectal cancer recognition.

Method: We proposed semi-supervised classifiers based on deep learning that provide pathological predictions at both patch-level and the level of whole slide image (WSI). First, we developed a semi-supervised deep learning framework based on the mean teacher method, to predict the cancer probability of an individual patch by utilizing patch-level data generated by dividing a WSI into many patches. Second, we developed a patient-level method utilizing a cluster-based and positive sensitivity strategy on WSIs to predict whether the WSI or the associated patient has cancer or not. We demonstrated the general utility of the semi-supervised learning method for colorectal cancer prediction utilizing a large data set (13,111 WSIs from 8,803 subjects) gathered from 13 centers across China, the United States and Germany. On this data set, we compared the performances of our proposed semi-supervised learning method with those from the prevailing supervised learning methods and six professional pathologists.

Results: Our results confirmed that semi-supervised learning model overperformed supervised learning models when a small portion of massive data was labeled, and performed as well as a supervised learning model when using massive labeled data. Specifically, when a small amount of training patches (~3,150) was labeled, the proposed semi-supervised learning model plus ~40,950 unlabeled patches performed better than the supervised learning model (AUC: 0.90 ± 0.06 vs. 0.84 ± 0.07,P value $= 0.02$). When more labeled training patches (~6,300) were available, the semi-supervised learning model plus ~37,800 unlabeled patches still performed significantly better than a supervised learning model (AUC: 0.98 ± 0.01vs. 0.92 ± 0.04, P value $= 0.0004$), and its performance had no significant difference compared with a supervised learning model trained on massive labeled patches (~44,100) (AUC: 0.98 ± 0.01 vs. 0.987 ± 0.01, P value $= 0.134$). Through extensive patient-level testing of 12,183 WSIs in 12 centers, we found no significant difference on patient-level diagnoses between the semi-supervised learning model (~6,300 labeled, ~37,800 unlabeled training patches) and a supervised learning model (~44,100 labeled training patches) (average AUC: 97.40% vs. 97.96%, P value $= 0.117$). Moreover, the diagnosis accuracy of the semi-supervised learning model was close to that of human pathologists (average AUC: 97.17% vs. 96.91%).

Conclusions: We reported that semi-supervised learning can achieve excellent performance at patch-level and patient-level diagnoses for colorectal cancer through a multi-center study. This finding is particularly useful since massive labeled data are usually not readily available. We demonstrated that our newly proposed semi-supervised learning method can accurately predict colorectal cancer that matched the average accuracy of pathologists. We thus suggested that semi-supervised learning has great potentials to build artificial intelligence (AI) platforms for medical

sciences and clinical practices including pathological diagnosis. These new platforms will dramatically reduce the cost and the number of labeled data required for training, which in turn will allow for broader adoptions of AI-empowered systems for cancer image analyses.

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

Introduction

Colorectal cancer (CRC) is the second most common cause of cancer death in Europe and America, with a lifetime incidence of up to 6%. By 2030, the global burden of CRC is expected to increase by 60%, causing more than 2.2 million new cases and 1.1 million deaths per year [1-2]. Pathological diagnosis is one of the most important and authoritative methods for diagnosing CRC [3-4]. Current pathological diagnosis of CRC requires a professional pathologist to visually examine digital full-scale whole slide images (WSI) of hematoxylin and eosin (H＆E) stained specimens. This process is typically labor intensive and time consuming. The challenges in WSI analysis stem from the complexity of WSI data 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 is effective yet of low cost by leveraging recent development in artificial intelligence (AI).

Deep learning provides an exciting opportunity to support and accelerate medical pathological analysis [6], and has been applied to assist diagnosis of various tumors including lung [7,8], breast [9,10], and skin cancers [11,12]. Progress has been made in applying deep learning to CRC studies including classification of cancer tissue [13], tumor cell detection and classification [14-15, 27], and outcome prediction [16-18]. For example, we have developed a recognition system for CRC WSI using a supervised learning method, which achieved one of highest diagnosis accuracies in general research area of cancer, our method even performed slightly better than some experienced pathologists [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 and high prediction accuracy in *patient level* pathology diagnosis. For example, on a small data set of 115 WSIs, a semi-supervised learning 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 evaluated and validated on *patient-level* data with large dataset from multiple centers to assess the generality of the utility of semi-supervised approach. 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 artificial intelligence system.

To fill this gap, we used a CRC dataset composed of 13,111 WSIs collected from 8,803 patient subjects from 13 centers to develop semi-supervised CRC recognition model at both patch level and the level of WSI and patient. We evaluated the performance of the semi-supervised approach by comparing its performance with that of prevailing supervised learning and also with that of professional pathologists. 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]. 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 collected from 13 medical centers. This large-scale evaluation showed that accurate recognition of CRC is feasible with a high degree of reliability even when the number of labeled data is limited.

(2) We found that semi-supervised model perform better than supervised model when only a small number of labeled patches（~3,150）is available (assume a large number of unlabeled patches (e.g., ~40,950) available for semi-supervised training, which is often the case in practice). We observed that when ~6,300 labeled and ~37,800 unlabeled patches are 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 learning model trained on ~6,300 labeled plus ~37,800 unlabeled patches can match the accuracy of professional pathologists. This result demonstrated the potential power of semi-supervised learning in an important medical application area. 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 tested our proposed semi-supervised method to CRC recognition (Figure 1, Table 1). Briefly, we divided each WSI into hundreds of patches. First, the cancerous probability was identified at the patch level. Next, the cancerous probability of WSI was inferred by using a clustering-based inference strategy. Finally, the prediction of patients with or without cancer was inferred based on the criteria conducive to positive diagnosis. We used several criteria including sensitivity, specificity, accuracy, and AUC (area under the curve) to evaluate the performance of various learning methods and pathologist diagnosis.

Semi-supervised vs supervised recognition at patch level

We aimed to evaluate two hypotheses at the patch level. First, we hypothesized that semi-supervised learning is better than supervised learning when only a small number (~thousands) of labeled patches available for both supervised and semi-supervised learning, and a large number of unlabeled patches is also available for semi-supervised learning. Second, we further hypothesized that there is no significant difference between semi-supervised learning using a few thousands of labeled patches (plus a larger number of unlabeled patches such as tens of thousands) and supervised learning using massive labeled patches (tens of thousands). To test these two hypotheses, we trained five models with different input data as described below.

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 plus ~40,950 unlabeled patches was superior to model-5%-SL (AUC: 0.90 ± 0.06 vs. 0.84 ± 0.07, P value $= 0.02$). With the availability of approximately 6,300 labeled and 37,800 unlabeled patches, the model-10%-SSL was also obviously better than model-10%-SL (AUC: 0.98 ± 0.01 vs. 0.92 ± 0.04, P value $= 0.0004$). These results indicated that when approximately 3,150 or 6,300 patches were labeled, the semi-supervised models were always better than the supervised models.

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 could not really find the pixels of cancer in the patches, while the pixels of cancer 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 evaluated three of 5 models using Dataset-PT. Model-5%-SSL and model-5%-SL were removed from subsequent experiments because they performed far worse than the other three models at the aforementioned patch-level comparison experiments.

As illustrated in Figure 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 $= 0.0022$), which indicated that the semi-supervised learning was significantly better than that of the supervised learning on patient-level prediction in the multi-centers scenario when the unlabeled ~37,800 training patches were included in the SSL. The average AUC of model-10%-SSL was slightly lower than, but comparable to, that of model-70%-SL with no significant difference (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 data centers (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 Figure 3 showed the cancer patches identified by model-10%-SSL and model-70%-SL were both highly matched, to 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) (Figure 4).

We ranked the average of six expert results, 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 experts (96.91%). The sensitivity of model-10%-SSL was 97.68%, ranked the 5th, showing an excellent detection ability of cancer (Supplementary Table 5).

Comparison with related studies

We compared our methods with 7 existing CRC methods with results shown in Supplementary Table 4. The first 6 of 7 methods had an AUC ranging from 0.904 to 0.99 based on supervised learning. Besides, the seventh used 86 subjects to develop a semi-supervised method, and used the test set of 7,180 patches of 50 WSIs from the same data source with the best accuracy of 0.938 confirming the potential of semi-supervised learning on patch-level. However, we showed the advantages of the semi-supervised method in 162,919 patches and 13,111 WSIs at both patch and patient levels from multiple centers, attesting to the general utility of the SSL model we developed.

Discussion

Pathological examination is an important cancer diagnosis method. However, accurately diagnosing pathological images requires years of training, leading to a global shortage of pathologists [2]. Almost all computer-assisted pathology diagnosis currently relies on massive labeled data with supervised learning approach, but labeled data is usually time-consuming and costly, due to one-by-one manual labeling process by medical experts. On the other hand, there exists a large amount of unlabeled data in clinics. 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 based on the mean teacher method for CRC diagnosis, and evaluated its performance using a large collection of WSIs across 13 medical centers in China, the United States and Germany, consisting of 13,111 WSIs from 8,803 patients. 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). In contrast, a supervised learning model trained by the same number (6,300) of labeled patches was difficult to achieve satisfactory results.

To demonstrate that semi-supervised learning can be used to achieve accurate CRC recognition, we conducted extensive testing of three models for patient level prediction on 12 centers (Dataset-PT). When the training data and testing data came from the same center (XH-Dataset-PT), the AUCs of model-70%-SL, model-10%-SSL and model-10%-SL were 99.16%, 98.41% and 96.44% respectively (Supplementary Table 2). 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%, maybe because 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 Figure 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 urgent need to develop semi-supervised learning methods, capitalizing on recent advances in deep learning.

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, so as to greatly reduce the tremendous cost of labeling required of professional pathologists.

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 with complex shapes, textures, and histological changes in nuclear staining. It is unclear whether existing semi-supervised methods can overcome the limitation of insufficient labeled pathological images. Our work confirmed that unlabeled data can improve CRC recognition and drastically reduced the number of required labeled patches. As demonstrated in our study, semi-supervised learning has great potentials to overcome the limitation 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. However, the annotation and labelling of data are often difficult, slow, and expensive. In this study, we applied a semi-supervised method for colorectal cancer recognition 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.

Acknowledgement

K.S.W was partially supported by the National Natural Science Foundation of China (#81673491) and the Natural Science Foundation of Hunan Province (#2015JJ2150). H.M.X was partially 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 partially supported by grants from 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, P.J., Willems, S.M., Wang, H., Madabhushi, A., Cruz-Roa, A., Gonzalez, F., Larsen, A.B., Vestergaard, J.S., Dahl, A.B., 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., van Ginneken, B., Karssemeijer, N., Litjens, G., van der Laak, J., the, C.C., Hermsen, M., Manson, Q.F., 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, O. J., De Raedt, S., Kleppe, A., Hveem, T. S., Liestøl, K., Maddison, J., 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]. Heller R, Stanley D, Yekutieli D, Rubin N, Benjamini Y. Cluster-based analysis of FMRI data. Neuroimage 2006, 33:599-608.

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

[27] Wei, J. W., Suriawinata, A. A. , Vaickus, L. J. , Ren, B. , Liu, X. , & Lisovsky, M. , 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.

Methodology

Datasets

Our dataset was composed of 13,111 WSIs collected from 13 sources, including 10 hospitals, a professional clinical laboratory (ACL), two public databases (Table 1). The data were divided into

four datasets (Dataset-PATT, Dataset-PAT, Dataset-PT, Dataset-HAC).

Digitization and annotation

In the 10 hospitals and ACL, the technicians randomly selected slides from archive library. The slides from 2010-2019 were scanned with a KF-PRO-005 scanner (KFBIO company, Ningbo City, China) at 20X magnification. The number of selected patients collected on the same day was limited to less than 50 to make sure the selected WSIs for this study were not unduly influenced by samples collected on any one single day.

Dataset-PATT, PAT and PT were reviewed by two experienced pathologists. If two experts disagreed with each other or with the previous diagnosis, the WSI would be excluded. Dataset-HAC was used for human-AI competition, and the review criteria were more rigorous. The WSIs were reviewed by five experienced pathologists and it was included if they reached an agreement.

From the 842 WSIs in Dataset-PATT, two pathologists manually selected some representative patches, each of which was 300 * 300 pixels in size and weakly labeled with either cancer or cancer-free. In order to ensure the diversity of the data, the number of selected patches on one positive WSI was not more than 100, and the numbers of patches with many and a few cancer cells were similar. Meanwhile, the number of various CRC subtypes were basically consistent with the subtype morbidity in the population. A total of 30,056 patches with cancer and 32,863 patches without cancer were obtained.

The Dataset-PATT was randomly divided into training set and testing set according to the proportions showed in Table 3, and the patches from the same subject would not be in different sets, to ensure independence of the different data sets. Meanwhile, the patches from 70% subjects were used as the training set, while the remaining 30% subjects were used as the testing set. Because training a deep learning model is time-consuming, for illustration, we repeated the process 8 times and calculated 75% confidence interval.

We trained five patch-level models (Table 3). In model-5%-SSL and model-10%-SSL, we used semi-supervised learning and keep labels for small proportions (i.e., 5% and 10%) of total patches (62,919) and masked label information for the remaining patches. In model-5%-SL, model-10%-SL and model-70%-SL, we used supervised training with 5%, 10%, 70% of the total 62,919 patches. The Dataset-PAT was used as an independent test.

Algorithm pipeline

Because the WSI is very large (>50,000 pixels), the patch-level models were trained to recognize cancerous probability, and all the patch-level results on a WSI were combined to infer the cancerous probability of the WSI. The flow chart is shown in Figure 1.

Patch-level models

We tested the known CNNs, such as VGG16, ResNET V1 and V2, Inception V1-V4, Mobilenet, etc., and found that InceptionV3 [26] achieved most consistent results on many datasets. Therefore, we used Inception V3 as the baseline model.

The patch-level models included supervised and semi-supervised versions. The input patch size was scaled to 299×299, the default input size of Inception V3. The top output layer was removed, and the output category was modified to two (cancer or non-cancer). The semi-supervised version was based on the mean teacher method [21], where two Inception V3 were trained, one as student and the other as teacher.

Network training at patch level

The Inception V3 adopted the pre-trained model on ImageNet database, and was deeply fine-tuned. We used the same preprocessing in protocols we used earlier [19]. All background patches without any cell tissue were removed. After data augmentation (image zoom, flip, color change), the grayscale of each pixel was normalized to [-1,1].

In semi-supervised learning, because of the imbalance between the labeled and unlabeled data, we maintained the same proportion of labeled and unlabeled patches in each mini-batch. The training cycle was 100 epochs, each epoch included 100 steps. L2 decay was used and deck weight was set to 0.0001. The teacher model was initialized with the student model, and the updated weight was set to 0.9. For unlabeled patches, unsupervised loss was obtained by calculating mean square from the pseudo labels given by the teacher model and the predicted label of the student model. For the labeled training data, the cross entropy of the predictive label and the real label was used to calculate the supervised loss. The weighted sum of the two loss was used to update the student model. The student model would update the weights in each step, but the teacher model used exponential moving average to update the weights after one epoch end.

Clustered-based WSI inference

Because the accuracy of patch-level models cannot be 100%, there were serious false positives in WSI predictions if patch level prediction is simply used to extrapolate the WSI cancerous status. Intuitively, because the tissues in WSI were continuous, the area with cancer should be distributed continuously and included several continuous patches. This intuition had been used to effectively control the false-positive of functional magnetic resonance images [25]. We designed a simple clustering-based inference method. If some continuous patches were identified as having cancer by patch-level model, the cancer may indeed exist on WSI. The cluster size of four patches was expected to best control the false-positive per our early study [19]. that is, the condition of continuously identifying 4 patches with cancer on WSI was used as the basis for determining the existence of cancer in WSI.

Patient-level diagnosis

Clinically, multiple WSIs may be obtained for one patient. The inference on patient level was based on positive sensitivity, that is, if all WSIs from the same patient were identified as negative (no cancer), then the patient was negative, otherwise the patient was positive.

For further information on the methodology, please refer to Supplementary Files A-C.

Table 1. Datasets used from multi-center data sources

Data source	Dataset Usage	Sample preparation	Examination type Radical surgery / Colonoscopy	Population	CRC		Non-CRC		Total	
					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: National Center for Tumor diseases, University Medical Center Mannheim, Heidelberg University, Germany, was downloaded at 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, b, c, d, e: About 3,150, 6,300, 44,100, 40,950, 37,800 patches, because the patches from 5%, 10%, 70%, 65% and 60% patients in Dataset-PATT are used, 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

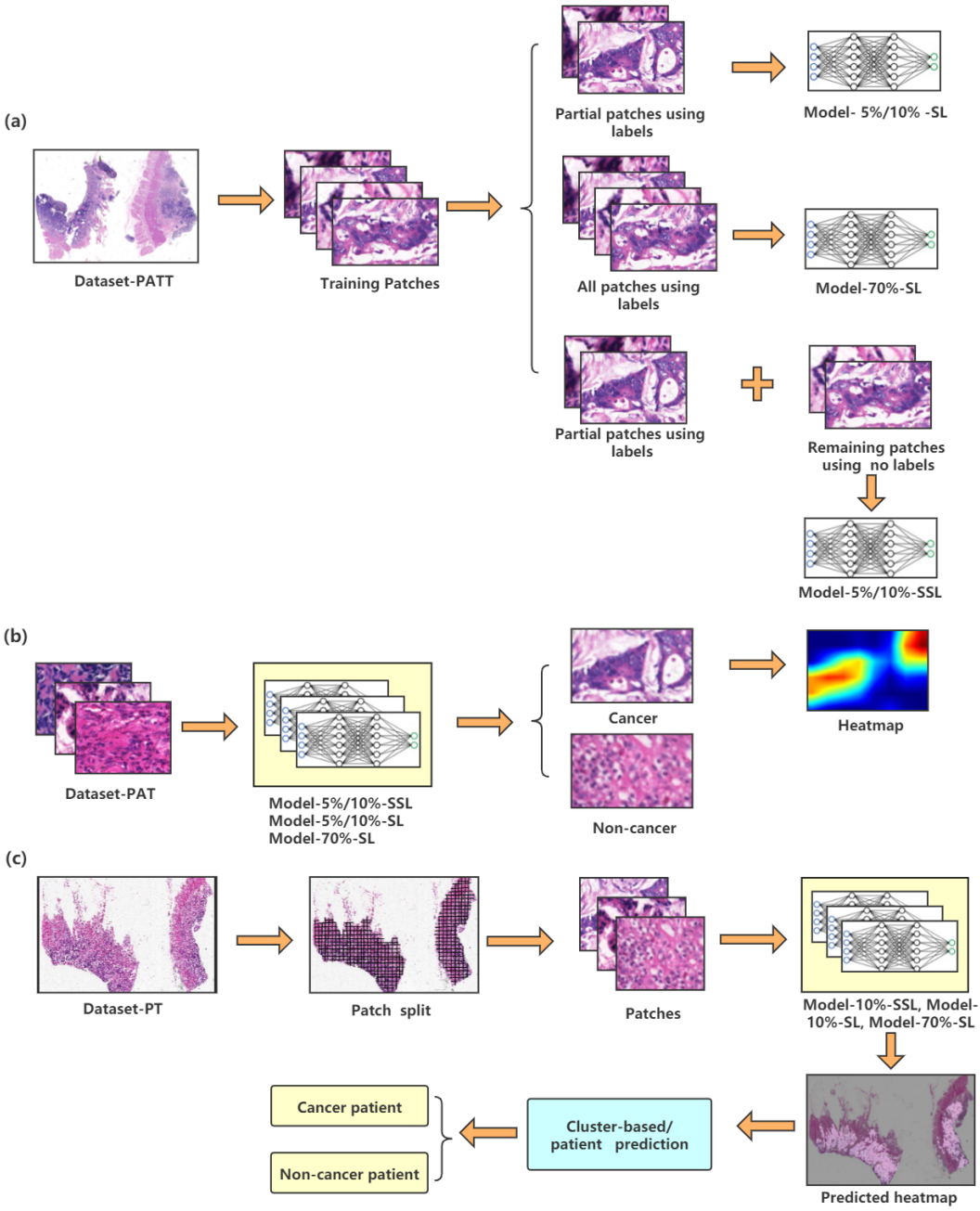
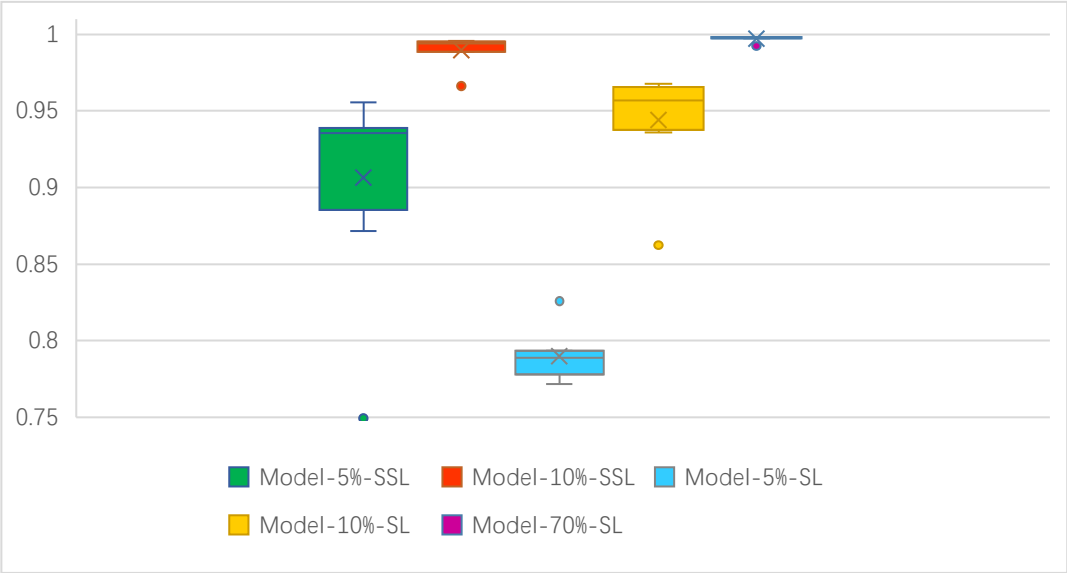
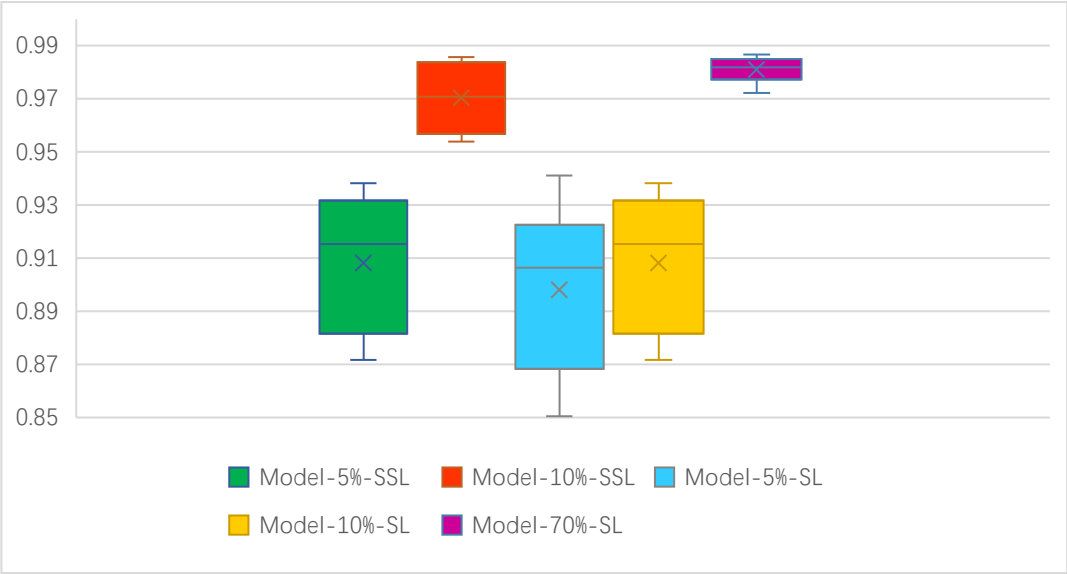


Figure 1. The flow chart of the study. (a) Semi-supervised and supervised training are performed on patches of the Dataset-PATT training set. (b) The patch-level test of five models on Dataset-PAT. (c) The patient-level test used Dataset-PT. The heatmap shows the cancer locations in WSI.

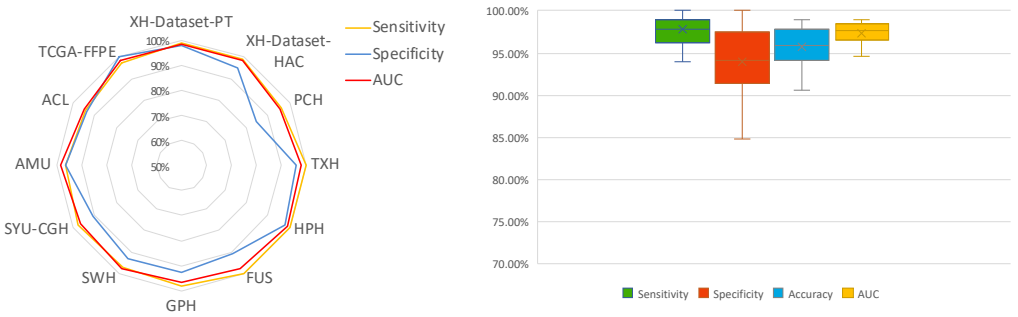


(a) Dataset-PATT (test)

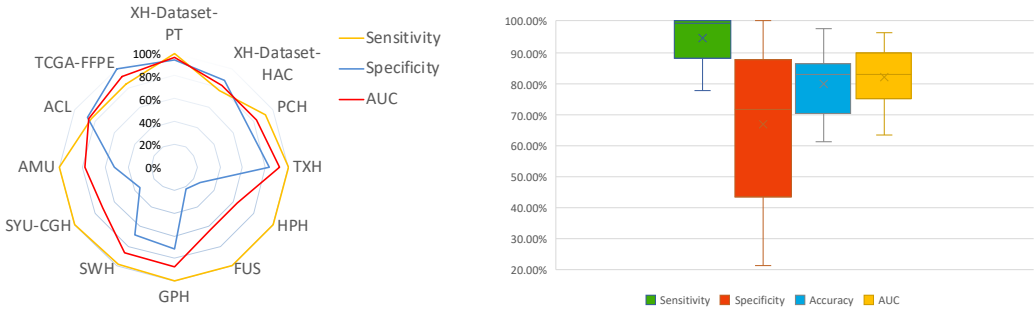


(b) Dataset-PAT

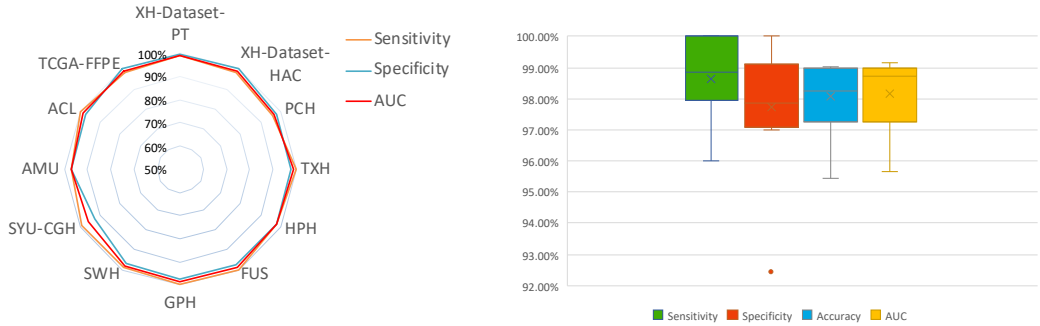
Figure 2. The AUC distribution of five models at patch level on two datasets.



(a) Model-10%-SSL



(b) Model-10%-SL



(c) Model-70%-SL

Figure 3. Patient-level comparison of model-10%-SSL, model-10%-SL and model-70%-SL 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. The horizontal bar in a boxplot indicates the median, while the cross indicates the mean of that distribution.

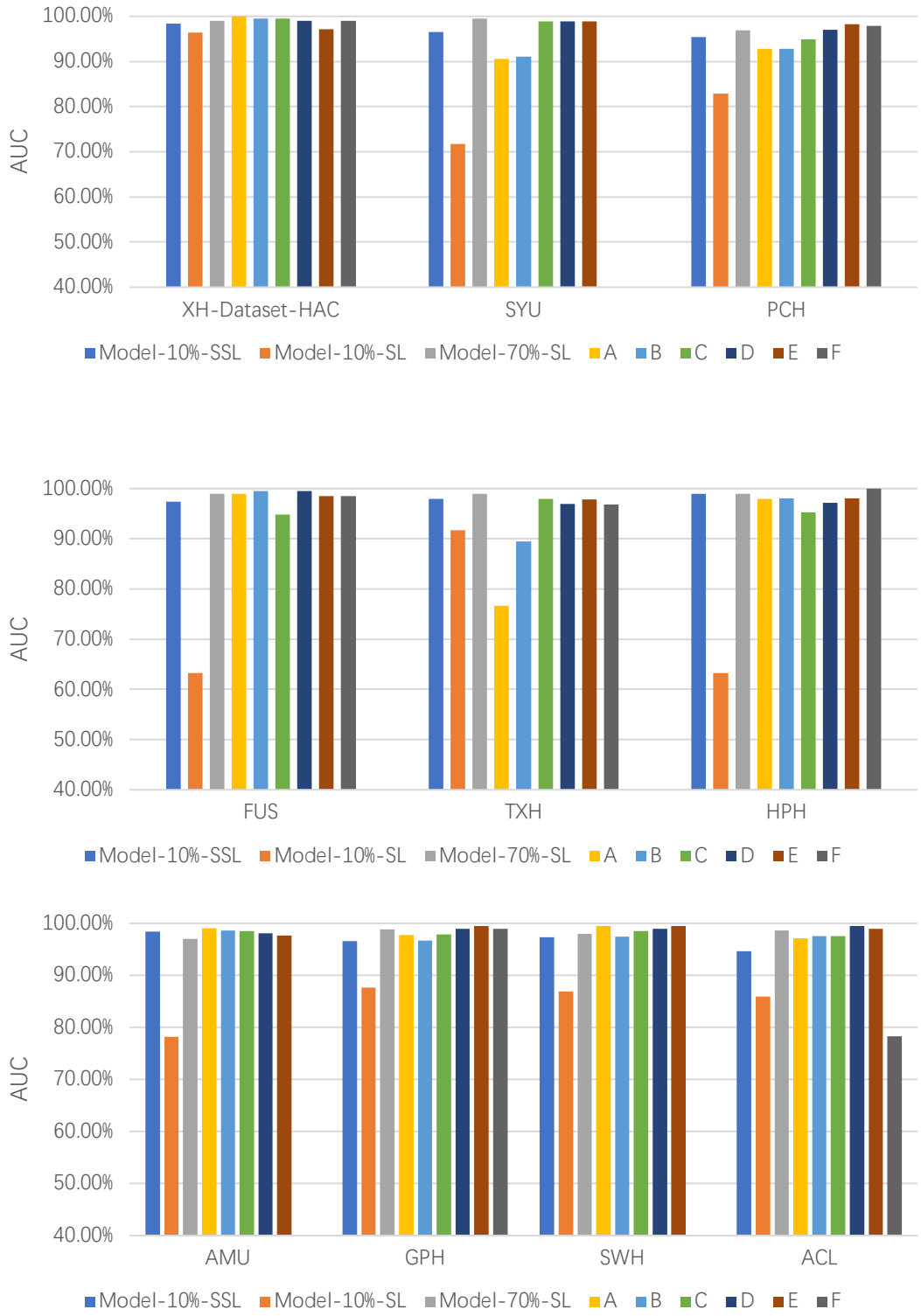


Figure 4. AUC comparison of in the Human-AI contest using Dataset-HAC. Colored lines indicate the AUCs achieved by three models and six pathologists (A-F).