Medical Image Analysis 67 (2021) 101854

Contents lists available at ScienceDirect

Medical Image Analysis

journal homepage: www.elsevier.com/locate/media

PAIP 2019: Liver cancer segmentation challenge☆

Yoo Jung Kim[a,1], Hyungjoon Jang[b,1], Kyoungbun Lee[c,1], Seongkeun Park[a], Sung-Gyu Min[c], Choyeon Hong[c], Jeong Hwan Park[d], Kanggeun Lee[b], Jisoo Kim[b], Wonjae Hong[b], Hyun Jung[e], Yanling Liu[e], Haran Rajkumar[f], Mahendra Khened[f], Ganapathy Krishnamurthi[f], Sen Yang[g], Xiyue Wang[h], Chang Hee Han[i], Jin Tae Kwak[i], Jianqiang Ma[j], Zhe Tang[j], Bahram Marami[k], Jack Zeineh[k], Zixu Zhao[l], Pheng-Ann Heng[l], Rüdiger Schmitz[m,n], Frederic Madesta[o,n], Thomas Rösch[m], Rene Werner[o,n], Jie Tian[p], Elodie Puybareau[q], Matteo Bovio[q], Xiufeng Zhang[r], Yifeng Zhu[s], Se Young Chun[b,*], Won-Ki Jeong[t,*], Peom Park[u,*], Jinwook Choi[a,*]

[a] Department of Biomedical Engineering, Seoul National University Hospital, Seoul, South Korea
[b] School of Electrical and Computer Engineering, Ulsan National Institute of Science and Technology, Ulsan, South Korea
[c] Department of Pathology, Seoul National University Hospital, Seoul, South Korea
[d] Department of Pathology, Seoul Metropolitan Government-Seoul National University Boramae Medical Center, Seoul, South Korea
[e] Frederick National Laboratory for Cancer Research, Frederick, Maryland, United States
[f] Department of Engineering Design, Indian Institute Of Technology Madras, Chennai, Tamil Nadu, India
[g] Sichuan University and Tencent AI Lab, Chengdu, Sichuan, China
[h] College of Computer Science, Sichuan University, Chengdu, Sichuan, China
[i] Department of Computer Science and Engineering, Sejong University, Seoul, South Korea
[j] Alibaba Group, China
[k] The Center for Computational and Systems Pathology, Icahn School of Medicine at Mount Sinai, New York, NY, United States
[l] Department of Computer Science and Engineering, The Chinese University of Hong Kong, Hong Kong
[m] Department for Interdisciplinary Endoscopy, University Medical Center Hamburg-Eppendorf, Hamburg, Germany
[n] DAISYlabs, Forschungszentrum Medizintechnik Hamburg, Hamburg, Germany
[o] Instutute of Computational Neuroscience, University Medical Center Hamburg-Eppendorf, Hamburg, Germany
[p] Shanghai JiaoTong University, Shanghai, China
[q] LRDE EPITA, France
[r] Tianjin Chengjian University, Tianjin Shi, China
[s] University of Maine, Orono, ME, United States
[t] Department of Computer Science and Engineering, College of Informatics, Korea University, Seoul, 02841, Korea
[u] HuminTec, Suwon, South Korea

ARTICLE INFO

Article history:
Received 31 December 2019
Revised 6 July 2020
Accepted 3 September 2020
Available online 8 October 2020

MSC:
41A05
41A10
65D05
65D17

ABSTRACT

Pathology Artificial Intelligence Platform (PAIP) is a free research platform in support of pathological artificial intelligence (AI). The main goal of the platform is to construct a high-quality pathology learning data set that will allow greater accessibility. The PAIP Liver Cancer Segmentation Challenge, organized in conjunction with the Medical Image Computing and Computer Assisted Intervention Society (MICCAI 2019), is the first image analysis challenge to apply PAIP datasets. The goal of the challenge was to evaluate new and existing algorithms for automated detection of liver cancer in whole-slide images (WSIs). Additionally, the PAIP of this year attempted to address potential future problems of AI applicability in

* Corresponding authors.
E-mail addresses: yjkim191@gmail.com (Y.J. Kim), jhj0110@unist.ac.kr (H. Jang), azi1003@snu.ac.kr (K. Lee), sychun@unist.ac.kr (S.Y. Chun), wkjeong@korea.ac.kr (W.-K. Jeong), ppark@ajou.ac.kr (P. Park), jinchoi@snu.ac.kr (J. Choi).

[1] Equal contribution.

https://doi.org/10.1016/j.media.2020.101854

Keywords:
Liver cancer
Tumor burden
Digital pathology
Challenge
Segmentation

clinical settings. In the challenge, participants were asked to use analytical data and statistical metrics to evaluate the performance of automated algorithms in two different tasks. The participants were given the two different tasks: Task 1 involved investigating Liver Cancer Segmentation and Task 2 involved investigating Viable Tumor Burden Estimation. There was a strong correlation between high performance of teams on both tasks, in which teams that performed well on Task 1 also performed well on Task 2. After evaluation, we summarized the top 11 team's algorithms. We then gave pathological implications on the easily predicted images for cancer segmentation and the challenging images for viable tumor burden estimation. Out of the 231 participants of the PAIP challenge datasets, a total of 64 were submitted from 28 team participants. The submitted algorithms predicted the automatic segmentation on the liver cancer with WSIs to an accuracy of a score estimation of 0.78. The PAIP challenge was created in an effort to combat the lack of research that has been done to address Liver cancer using digital pathology. It remains unclear of how the applicability of AI algorithms created during the challenge can affect clinical diagnoses. However, the results of this dataset and evaluation metric provided has the potential to aid the development and benchmarking of cancer diagnosis and segmentation.

1. Introduction

Artificial intelligence (AI) is becoming increasingly prevalent in society, with recent applications emerging in the healthcare sector. Pathology is one of the most rapidly growing areas where in deep learning is used for analyzing the medical images (Litjens et al. (2017)). In one study, a pathologist and an AI-based algorithm were given images of lymph node cells and were assigned the task of determining whether or not the cells were cancerous. It was found that the results were more accurate when they were derived using both of the sources than when either of the sources was used, and an error rate of 0.5% was achieved (M. Holden and Smith (2016)).

South Korean government has been promoting research in AI by increasing research funding and making large investments in the AI industry. Since 2018, the Ministry of Health and Welfare in Korea has been supporting three AI platform projects aimed at paving the way for the rapid application of AI in clinical diagnoses by focusing on building a training dataset for AI researchers.

The Pathology AI Platform (PAIP) (PAIP (2019)) is a free research support platform of pathology-related AI. The main goal of the platform is to construct high-quality pathology learning datasets, which are provided by the Seoul National University Hospital, Seoul National University Bundang Hospital, and SMG-SNU Boramae Medical Center. These datasets are collected over three years (2018–2021) and include over 3000 whole-slide images for six diagnostic cancers. The PAIP Liver Cancer Segmentation Challenge is the first image analysis challenge to apply the PAIP datasets. In this challenge, participants were tasked with using analytical data and statistical metrics to evaluate the performances of automated algorithms in determining liver cancer segmentation or viable tumor burden (TB) estimation. PAIP has submitted the "PAIP 2019 Challenge" as a part of the Medical Image Computing and Computer Assisted Intervention Society (MICCAI) 2019 Grand Challenge for Pathology, which aims to generate new insights into and identify common themes for future cancer research. This was an effort to counter the lack of challenges that have been done to address Hepatocellular carcinoma (HCC) diagnosis using digital pathology. Despite changes in the population demographics, the prevalence of liver cancer, specifically, has not yet been addressed by medical image challenges. Currently, only computed tomography scans are available for use in research.

1.1. Related work

1.1.1. Liver cancer / tumor burden estimation

The liver is a multi-functional organ in the abdomen that plays a major role in the synthesis and detoxification of biomaterials in the bloodstream from the Gastrointestinal tract to the heart. HCC represents the most common histology of primary liver cancer and is originated from hepatocytes, which account for 60% of the liver volume. HCC is the most commonly found cancer in liver cirrhosis or chronic liver diseases (Akbar et al. (2019)) associated with alcohol, viruses, or metabolic diseases such as obesity or diabetes mellitus. Instances have increased by 75% between 1990 and 2015 due to changing age structures and population growth, which are the leading factors that contribute to cancer mortality globally (Longo (2019)). The tumor is composed of heterogeneous cellular components. Neoplastic hepatocytes are the main cells involved in HCC, while the remainder is composed of neoplastic tumor cells in the stromal tissue, blood vessels that feed on cells, and inflammatory cells that infiltrate the individual tumor cells. The latter represents the so-called "tumor environment," which plays a major role in tumor growth, response to treatment, and patient prognosis (Seok et al. (2012)) (Song et al. (2015)). Other components of the tumor mass involve a secondary change that may result from pretreatment, tumor characteristics, or tissue artifacts due to tissue manipulation. Ischemic necrosis, tumor necrosis, hemorrhages, and cystic changes are examples of such secondary changes.

The tumor burden was used in radiology to evaluate the therapeutic efficacy of treatment following treatment. In pathology, the tumor burden is an important parameter used to estimate treatment response and the performance of molecular testing. In classical pathology, there is no protocol for TB assessment. Recently, a detailed pathology protocol to determine the residual cancer burden in a breast cancer specimen was published by the MD Anderson Cancer Center using a number of tissue sections and manually-estimated cellularity (A.C. (2019)).

Necrosis in HCC is usually induced by pretreatment embolization or radiofrequency ablation before the operation. The protocol of the College of American Pathologists recommends reporting a treatment in terms of the percentage of necrosis that occurs (Sanjay Kakar and Washington (2017)). Although the extent of necrosis in a pathologic evaluation can be valuable for correlating with radiologic images (Yao et al. (2008)), the direct prognostic relevance in terms of the outcome of the patient is not known (Cotoi et al. (2012)).

1.1.2. Digital pathology image analysis

Digital pathology (DP) is the process by which histology slides are digitalized to produce high-resolution images (Janowczyk and Madabhushi (2016)). There have been studies on how to acquire, process, and interpret the digitalized pathological slide images (Gurcan et al. (2009)). Technological advances have created a paradigm shift from digitalized slide images towards digital pathology. The combination of the whole-slide imaging (WSI) technology and big data analytics has enabled access to unprecedented details about data from the subcellular to the tissue level (Bhargava and Madabhushi (2016)). The variety of image analysis tasks in the context of DP includes detection and counting (e.g., mitotic events), segmentation (e.g., nuclei), and tissue classification (e.g., cancerous vs. non-cancerous) (Janowczyk and Madabhushi (2016)). Increasing usage of TB in pathology combined with advances in digital scanning technology has led to increased attempts to measure TB using automated algorithms with higher accuracy and precision. S. Akbar et al. compared two methods of automated TB measurement in breast cancer slides: a hand-engineered feature approach using traditional image processing techniques, and a "deep convolutional neural networks" approach in which image features were automatically extracted. The two automated methods showed a strong correlation with pathologists' assessments (Akbar et al. (2019)).

1.2. Medical image analysis challenge

The PAIP 2019 challenge attempts to suggest solutions to important problems of AI applicability in clinical use. Firstly, an environment was created using digital pathology, which most pathologists utilize to diagnose cancer. This challenge provides the contestants with whole-slide images rather than small image tiles. These whole-slide images create technical hurdles for image analysis, but their use was an important component in the application of AI for clinical purposes. Secondly, a task was designed for viable tumor burden estimation. The viable tumor burden is the ratio of the viable tumor area to the whole area of the tumor. The need for the evaluation of the viable tumor burden has increased based on the assessment of the response rates toward chemo-radiotherapy or the proportion of tumor cells determined via genetic testing using tissue samples. Traditional pathologists either use a semi-quantitative grading system to determine the residual tumor burden or report the portion of necrosis, thereby indirectly indicating the viable tumor burden. The main problem with the evaluation of the viable tumor burden is the uncertainty concerning the extent of the tumor as a whole relative to the extent of the viable tumor cells.

The determination of tumor cells and necrotic cells using whole-slide images has been achieved by experienced pathologists. This process is labor-intensive, in addition to being unsalable to translational and clinical research studies that involve hundreds of resected specimens. Machine learning and AI-based diagnoses have improved the process of tumor diagnosis and enabled the possibility for quantitative studies of the mechanisms and progression of the disease. In this paper, we will present an overview of the PAIP 2019 challenge, addresses the prediction results, and discuss some of the research problems addressed in the medical image analysis challenge.

2. Challenge description

2.1. Organization

The goal of this year's challenge was to evaluate new and existing algorithms for the automated detection of liver cancer using whole-slide images. The PAIP organizers designed two tasks for the PAIP 2019 challenge. In the first task, contestants were invited to develop an algorithm to detect and segment areas of carcinogenic cells in terms of tumor viability. In the second task, participants were tasked with developing algorithms to assess and calculate the area of the tumor burden. Two leaderboards were established to evaluate the performances of the algorithms. Contestants were invited to participate in both tasks or the task that best aligned with their interests.

PAIP 2019 was hosted on Grand Challenge, which provides a user-friendly interface that allows for an efficient platform set-up. It is one of the preferred platforms for use in challenges in medical imaging analysis (Aresta et al. (2019)). Participants in the challenge were instructed to register on Grand Challenge to access the dataset, submit their algorithms, and view the evaluated results of their submissions. The Korean government has strict policies and regulations concerning the sharing of medical data. To protect the right to privacy and prevent the state's intervention into the private lives of citizens, the Personal Information Protection Act was legislated in 2011 (MOIS (2011)). All candidates had to read and consent to the "Data Use and Confidentiality Agreement" to confirm their eligibility and to access the dataset. Participants were required to provide their name, affiliation, address, email address, and handwritten signature. Their information was then manually validated by an event organizer. Once approved, participants received credentials (username and password) to access the PAIP platform, which contained the dataset and a download link.

This allowed the organizers to screen and differentiate between validated users and anonymous participants. The ground-truth information for both tasks was given to the participants for the training dataset. However, for the validation and test dataset, the ground-truth information was reserved for the challenge committee and used to evaluate the performance of the AI learning model of each participant. The PAIP 2019 challenge provided original whole-slide images, XML annotations made by pathologists, ground-truth binary pixel masks generated from the XML annotations for both the whole tumor area and viable tumor area, and the viable tumor burden calculated from the binary pixel masks.

The challenge website was launched with the provision of the training datasets. The training datasets were divided into two groups: a smaller group containing 80% of the training data and another group used for validation, which contained the remaining 20% of the training data. The training data was provided twice: on April 15th and May 20th. The validation dataset was released on August 12th and participants could upload their submissions on the challenge website. We established a daily-submission limit to prevent the overfitting of the model. Once a participant submitted their results, they were automatically evaluated and published on the results page. The test dataset (which is independent of the training and validation sets) was released on September 2, 2019. The test submission page was made available for seven days and a limit of seven submissions per participant was imposed. The final scores were not published on the leaderboard. Instead, PAIP invited the top 10 contestants to the MICCAI 2019 conference to present their results. The challenge workshop was held in Shenzhen, China on October 17th. The leaderboard was published on October 23rd and the submission for the test dataset was reopened.

2.2. Dataset

The dataset contained 100 WSI that were used for training (50), validation (10), and testing (40). The WSIs for training and validation have two-layers of annotation for the viable tumor area and whole tumor area. The data annotation was made by two expert pathologists. One pathologist with 11 years of experience in herpetology drew the boundaries of the whole-tumor areas and viable tumor areas. The second pathologist confirmed the annotation by screening for any missed or over-estimated tumor areas. The WSI

for testing were presented without annotation and information of the presence of a tumor.

2.2.1. Selection of dataset and preparation of WSI

All datasets were selected from the pathology archives of the Seoul National University Hospital obtained between 2000 and 2018. The training and validation datasets were composed of resection specimens of HCC, and the testing set was composed of 31 resected HCCs and 9 biopsy specimens. Slides presenting both tumor and non-tumor areas were preferentially selected for this challenge excluding biopsy samples. WSIs were acquired using an Aperio AT scanner (Leica Biosystems Imaging Inc. USA) at an x200 resolution with a file format of.svs, while the annotation files were.xml.

2.2.2. Annotation of viable tumor and whole tumor area

A pathologist constructed pixel-based annotations using an Aperio Image Scope (V12.4.0.5043, Leica Biosystems Pathology Imaging, USA). The viable tumor area was distinguished using a closed circle for one continuous tumor area under x200 magnification to minimize the intervening effects of the non-tumor cell components. Intra-tumoral necrosis, hemorrhage, and fibrous stroma, which were visible at as low as an x50 magnification, were removed using a negative pen tool. The whole tumor area indicates the boundary between the non-tumorous hepatic lobules and the viable tumor boundary, including peritumoral fibrosis, capsules, and inflammation. Because the histology of the whole tumor area was not specified as tumor cells and was a concept of territory, the whole tumor area was annotated at low power (x12.5) and refined at high power (x200) to exclude the normal structures.

2.2.3. Data preparation and release

The ground-truth area maps manually generated by pathologists were further converted into the final image format for release. The initial ground-truth XML file consisted of a list of vertices representing the closed polygons, which were hand-drawn by pathologists. The provided XML file contained two sets of area maps: one for the viable tumor area and the other for the whole tumor area. Each area map (denoted using polygons) was converted to a binary map (image) using a polygon scan-conversion algorithm at the highest WSI resolution (level 0). Because the tumor area map may have included non-tissue areas, further processing is required to generate more accurate area maps. We used intensity-based thresholding to generate a tissue mask (pixels having an intensity of less than RGB (235, 210, 235) are considered tissue) and removed fragments smaller than 10 pixels in size. The extracted tissue mask was then combined with the binary tumor area maps to generate the final ground-truth area maps. We allowed the participants to further improve the tissue mask using the provided raw XML annotation data.

Each slide in the training set includes an anonymized WSI in a `.svs` format, an original XML annotation for reference, a binary pixel mask for the ground-truth viable tumor area, and a binary pixel mask for the ground-truth whole tumor area. In addition, we released a single `.csv` file listing the viable tumor burden ratio of every slide in the training dataset, calculated using the ratio between the viable and whole tumor mask areas, which served as the ground-truth for Task 2.

2.3. Leaderboard management

We managed the leaderboard built on the `evalutils` scoring system provided by the Grand-Challenge platform. By default, the submission is automatically graded and the results (including the score and ranking) are posted to the leaderboard. During the validation submission phase, the participants were allowed to submit their results up to 10 times per day to tune their models and post-processing parameters without the ground-truth. During the test submission phase, which lasted for seven days, the participants were allowed to make only a single submission per 24-h period. The test submissions were graded internally and the results were not posted to the leaderboard to prevent the misuse of the system by the participants to overfit their models to the test dataset. The best result among seven submissions (at most) was used in the final rankings.

2.4. Metrics and evaluation

Two performance evaluation metrics were developed. One is based on the widely-used Jaccard index (for Task 1) and the other is the absolute difference of values (for Task 2).

2.4.1. Task 1: Viable tumor area segmentation

The performance of Task 1 was evaluated by measuring the pixel-wise similarity between the segmentation result (binary mask for viable tumor region prediction) and the ground-truth mask. For this, we used the clipped Jaccard index (Eq. (2)) to measure per-slice segmentation accuracy, which is the Jaccard index (Eq. (1)) clipped using a threshold value of 0.65 to penalize inaccurate results. For example, selecting the entire image area would result in a non-zero Jaccard index, which should not be counted as a meaningful result.

$$J(A,B) = \frac{|A \bigcap B|}{|A \bigcup B|} \tag{1}$$

$$\hat{J}_i = \begin{cases} J(M^i_{result}, M^i_{GT}), & \text{if } J(M^i_{result}, M^i_{GT}) >= 0.65 \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

where M^i_{result} is the segmentation result and M^i_{GT} is the ground-truth area map for the i-th WSI. Then, the final aggregated score for Task 1 is calculated as the average of the per-slice clipped Jaccard indices as follows:

$$\text{Score}_{Task1} = \frac{1}{N}\sum_{i=1}^{N}\hat{J}_i \tag{3}$$

where N is the number of WSIs in the testing dataset.

During the validation submission phase, we allowed the participants to only access the aggregated score and did not reveal the per-slice Jaccard index to prevent the misuse of the leaderboard.

2.4.2. Task 2: viable tumor burden ratio estimation

Task 2 involved the estimation of the viable tumor burden, which is the ratio between the viable and the whole tumor areas. The participants were asked to submit a list of tumor burden values in a `.csv` format. To evaluate the performances of Task 2, the absolute difference of the tumor burden values for a given slice i is calculated as follows:

$$TB_Diff_i = 1.0 - (|TB^i_{result} - TB^i_{GT}|)/100.0 \tag{4}$$

where TB^i_{result} is the submitted tumor burden and TB^i_{GT} is the ground truth tumor burden for the i-th WSI. Note that the tumor burden ranges between 0 and 100.

Unlike Task 1, only the numbers were submitted without the whole tumor area masks for Task 2. To avoid submitting arbitrary numbers without actually estimating the tumor regions, we added a weighting value considering the Task 1 score, giving more confidence if the Task 1 score was higher. The final aggregated Task 2 score was computed as an average of the per-slice weighted burden differences as follows:

$$Score_{Task2} = \frac{1}{N}\sum_{i=1}^{N}\hat{J}_i * TB_Diff_i \tag{5}$$

Table 1
Background of selected whole slide images.

Features		Training [50]	Validation [10]	Testing [40]	Total [100]
Year	~ 2000	2	0	0	2
	2001 ~ 2010	46	5	13	64
	2010 ~ 2018	2	5	27	34
Tissue type	Resection	50	10	31	91
	Biopsy	0	0	9	9
Grade (Edmonson-steiner)	Grade 1	7	0	3	10
	Grade 2	23	6	15	44
	Grade 3	20	4	22	46

where N is the number of WSIs in the testing dataset.

3. Methods

3.1. Summary of submitted methods

Since our PAIP challenge website was opened, 231 participants have registered to download the data. Among them, 28 teams submitted the final results during the testing phase. In this section, we summarize the methods from the top 11 teams (see Table 2 and Table 3) who submitted their extended abstracts and gave a presentation at the PAIP 2019 workshop at MICCAI 2019.

All 11 teams used similar preprocessing methods (e.g., decomposing the WSI into small patches, collecting patches in the tissue regions, applying color normalization and data augmentations, etc). All of these teams used deep convolutional neural networks or variants thereof (mostly encoder-decoder architecture such as U-Net (Ronneberger et al. (2015))) for tumor area segmentation. Multi-scale techniques and ensemble methods were also commonly used by many of the top-ranking teams. Detailed descriptions of the methods used by the top five teams are provided below.

3.2. Jung et al. (1st place, newhyun00)

Jung et al. ranked first in both Task 1 and 2. They used an ImageNet pre-trained EfficientNet-B4 (Tan and Le (2019)) as a backbone encoder structure, which has shown superior performance with high efficiency in resource usage. In addition, they employed a U-Net architecture (UNet++ level 5, Zhou et al. (2018)) for pixel-level segmentation. Unlike a common approach of setting the learning rate over cross-validation, they used fast.ai (2019) to find the best learning rate for the Rectified Adam(Liu et al. (2019)) optimizer. In the data preprocessing step, the authors merged the viable and whole tumor areas into one label image containing three classes (viable tumor, whole tumor excluding the viable tumor area, and background) based on the observation that the viable and whole tumor areas are related to each other. The proposed method employed an ensemble of nine networks for Task 1, which was trained with various magnification and learning rates using a simple majority voting method. For Task 2, the average of the nine viable tumor burden ratios was calculated using the results of Task 1. The results showed that the leveraging of various optical magnifications is one of the important factors contributing to the superior performance of their algorithm.

3.3. Yang et al. (2nd place, team sen)

Yang et al. ranked second in Task 1. They proposed multi-task learning over a modified U-Net, which adopts the SE-Resnext101 (Hu et al. (2018)) module over the encoder portion, and the SK block (Li et al. (2019)) is applied in the decoder portion. The authors applied the latent features of the encoder as an input to the patch-wise classification network. Furthermore, they extracted the latent features in the middle of the decoder as an input to the pixel-wise segmentation. Finally, concatenating all of the latent features of the decoder was used to generate the final segmentation results. The novel aspect of their proposed method involves the performance of patch-level classification and pixel-level segmentation together using a shared encoder-decoder network and leveraging state-of-the-art deep network modules (SE-Res and SK modules) with deep supervision (i.e., supervision on the intermediate layers). The Otsu thresholding method (Otsu (1979)) was used to remove the non-tissue region when processing the training data.

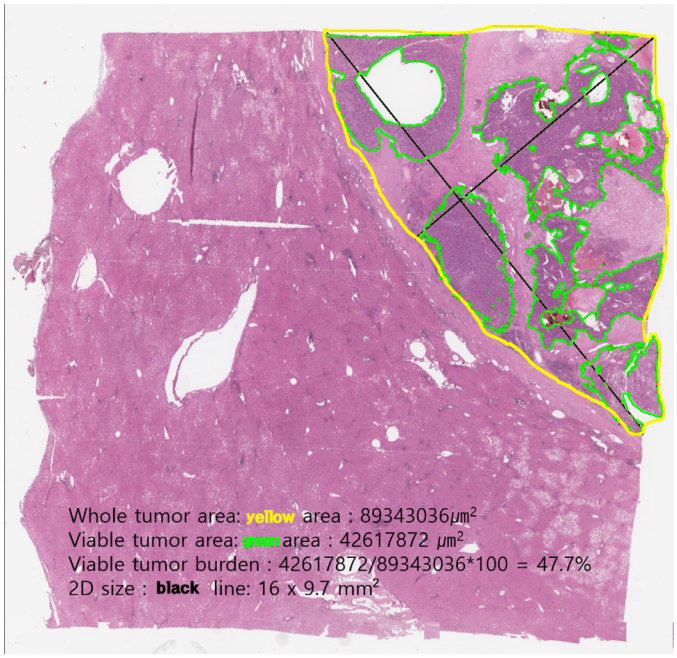

Fig. 1. Example of a viable tumor and a whole tumor area annotation.

3.4. Rajkumar et al. (3rd place, team MIRL-IITM)

Rajkumar et al. proposed an ensemble prediction method using three different deep neural networks: DenseNet-121 (Huang et al. (2017)), InceptionResNetV2 (Szegedy et al. (2017)), and DeepLabV3+ (Chen et al. (2018)). The first two networks use the ImageNet pre-trained backbones for the encoder and a conventional U-Net for the decoder, while the third network is trained from scratch. Their final results represent an average of the three different results obtained from these three networks. Various data

Table 2
Overview of the top[2] competitors' methods.

Rank (Task1 / Task2)	Team (Affiliation)	Model Architecture	Pre-trained	Learning rate	Optimizer	Loss func.	Framework	HW Devices	Epochs	Batch size
1 / 1	newhyun00 (Frederick National Lab)	UNet+ with EfficientNet-B4 backbone[3]	ImageNet (backbone)	LR finder from fast.ai (6.8e-5 @ 200X, 5.8e-5 @ 150X, 1.0e-4 @ 100X) Cosine annealing	RAdam	Categorical CE Jaccard	Keras	GTX 1080 Ti * 4ea	24	12
2 / -	Sen (Sichuan University)	UNet+ with classification outputs, SE-ResNeXt101 backbone, SK block decoders	ImageNet	3e-4 (epoch $\leq$ 20), 1e-4 (20 $<$epoch $\leq$ 30), 1e-5 (epoch $>$30)	Adam	CE Dice	Pytorch	Xeon E5-2680v4 Tesla P40s * 2ea	50	32
3 / 2	MIRL-IITM (Indian Institute of Technology Madras)	UNet with DenseNet121 backbone, UNet with InceptionResNetV2 backbone, DeepLabV3+[4]	ImageNet (UNet backbones) None (DeepLabV3+)	1e-5 (UNet: epoch $\leq$ 2, backbones frozen), 5e-6 (DeepLabV3+: epoch $\leq$ 2), 2e-4 (epoch $>$2)	Adam	CE Dice	Tensorflow	Core i7-4930K TITAN V	15 ~30	32
4 / 4	Damo AIC (Alibaba Inc.)	ResNet34/101 ASPP UNet (fully customized)	None	1e-3	Adam	CE Dice Focal loss	Pytorch	-	100	4
5 / 3	QuIIL (Sejong University)	DenseNet (Classification), Dense-UNet (Segmentation)[5]	None	1e-4	Adam	CE (Both), Focal loss (Dense-UNet)	-	TITAN Xp	60	10 (DenseNet), 20 (Dense-UNet)
6 / 6	CUHK-Med (Insight Medical Tech Inc. The Chinese Univ. of Hong Kong)	UNet with attention block and ResNet101 backbone, PFA-ScanNet	ImageNet (PFA-ScanNet)	1e-3 (UNet), 1e-4 (PFA-ScanNet) Cosine decay	Adam (UNet), SGD (PFA-ScanNet)	CE Dice	Keras	TITAN Xp	3000	-
7 / 7	DAISYlab@UKE (Universitäts-klinikum Hamburg-Eppendorf)	UNet-based multi-scale FCN with ResNet backbone (msYI-Net)[6]	ImageNet (ResNet backbone)	1e-3 (decayed by 0.5 for every 38.4K iterations)	Adam SGD	CE	Pytorch	Quadro P6000	$>$120	14
8 / 5	COSYPath (Icahn School Medicine at Mt. Sinai)	SegNet(1) (viable vs. others), SegNet(2) (viable vs. non-viable vs. others)	ImageNet	1e-4 (SegNet(1)), 1e-5 (SegNet(2))	Adam SGD	CE	Pytorch	GTX 1080 TITAN Xp	$<$100	8
9 / 11	LRDE (EPITA - LRDE)	VGG16 with decoder[7]	ImageNet	-	Adam	Categorical CE	-	-	50	8
10 / 8	Sig-IPath (12sigma technologies Inc.)	UNet	None	0.01 ReduceL-ROnPlateau	SGD	Dice	Pytorch	-	$<$100	32 (200X) 1 (12.5X)
14 / 9	blackbear (University of Maine Tianjin Chengjian University)	DeepLabV3+ UNet++ with ResNet152 backbone, UNet++ with InceptionResNetV2 backbone	ImageNet	1e-2 (divided by 10 after 3 or 4 epochs)	Adam	CE Dice	-	Tesla V100 * 1ea Tesla K80 * 4ea	$<$14	16 for V100 4 for K80

[3] There were top ten contestants on the leaderboard in each task, but only eligible teams were invited for the MICCAI 2019 conference. We decided to analyze the algorithms of the teams who have submitted the extended abstract and have presented their work at the conference. Therefore, the methodologies of the top nine teams for each task are summarized in this table. You may find the whole leaderboard on the challenge website.: https://paip2019.grand-challenge.org/ [3] https://github.com/newhyun00/paip2019_code [4] https://github.com/koriavinash1/DigitalHistoPath [5] https://github.com/ChangHeeHAN/PAIP2019 [6] https://github.com/RSchmitzHH/multiscale and https://arxiv.org/abs/1909.10726 (Schmitz et al. (2019)) [7] https://www.lrde.epita.fr/wiki/Challenges_codes (Puybareau et al. (2018)).

Table 3
Data processing and training details of the top competitors' methods.

Rank (Task1 / Task2)	Zoom level	Tissue detection	Stain norm	Num of patches	Input size	Augmentation	Etc.
1 / 1	200X 150X 100X	Threshold by average pixel values	X	175K[8]	512 * 512	Blurring, Shifting, Scaling, Rotating, Elastic transform (albumentations)	• Average outputs from 200X, 150X, 100X • Majority voting from 9 versions of output • Stochastic weight averaging
2 / -	200X	Otsu (Otsu (1979))	X	>1M	256 * 256	Flipping, Translating, Rotating, Color jittering	-
3 / 2	200X	Otsu + Closing / Opening / Dilating	X	>200K[9]	256 * 256 (training) 1024 * 1024 (testing)	Flipping, Rotating, Blurring, HSV, Contrast Center perturbation	• Whole tumor predictions were generated from convex hull of viable tumor predictions
4 / 4	200X 50X 12.5X	-	X	1K[10]	1024 * 1024 (200X model) 2048 * 2048 (50X model) 4096 * 4096 (12.5X model)	Rotating	• Average outputs from each model
5 / 3	200X	As the challeng website showed (RGB threshold + morphology)	X	53K (Classification) 22K (Segmen tation)[11]	1024 * 1024 (Classification) 512 * 512 (Segmentation)	Scaling, Shifting Rotating, Shearing Blurring, Color jittering	• Cascaded inference: viable tumor predictions were conducted only under whole tumor prediction area
6 / 6	200X	Otsu	X	5K	512 * 512 (UNet) 692 * 692 (training PFA-ScanNet) 2708 * 2708 (testing PFA-ScanNet)	HSV, Flipping, Scaling Rotating, Cropping Color jittering	• Aggregate two outputs to predict viable tumor burden: (1) UNet for viable tumor prediction (2) PFA-ScanNet for whole tumor prediction
7 / 7	200X 50X 12.5X	Color thresh-old + connected components	O (Reinhard et al. (2001))	Variable (Extract on-the-fly)[12]	512 * 512 (200X) 572 * 572 (50X) 512 * 512 (12.5X)	Rotatng, Flipping Brightness, Contrast	• 5-fold CV splits ensemble • Multi-scale multi-encoder model with a single decoder using all scales at the same time
8 / 5	20X	As the challenge website showed (RGB threshold + morphology)	O (Bejnordi et al. (2015))	-	500 * 500	HSV	• Aggregate outputs from two models trained with: (1) Viable tumor vs. Others (2) Viable tumor vs. Non-viable tumor vs. Others
9 / 11	50X	-	X	-	300 * 300	None	-
10 / 8	200X 12.5X	Empirical combinations of OpenCV threshold functions	X	1M (200X) 5K (12.5X)	512 * 512 (200X) 2048 * 2048 (12.5X)	Rotating, Flipping Affine transform Scaling, Lighting	• Multiply outputs from two models trained with: (1) 200X; (2) 12.5X • Negative mining strategy
14 / 9	200X	Otsu	X	391K (training) 45K (validation)	512 * 512 (training) 2048 * 2048 (testing)	HSV, Contrast, Flipping Rotating, Scaling Shearing, Cropping Elastic deforming Color jittering	• Exhaustive searching to find an optimal threshold value.

[9] The same number of patches were randomly sampled from three classes: Viable tumor, Non-viable tumor, and Background area. [9] The same number of patches were randomly sampled from two classes: Tumor and Non-tumor area. [10] Patches were randomly sampled after segmenting areas from WSIs into three: Foreground, Background, and Edge region. [11] For classification models, only patches with >50% tissue area were selected. Among them, patches with >80% tumor area were defined as tumor patches, while those with <20% as non-tumor patches. For segmentation models, patches were collected only from the area with >80% tissue area (i.e. tumor area). [12] The same number of patches were randomly sampled from four classes: Background, Overall tissue, Tumor, and Viable tumor.

augmentation techniques were used, such as flipping, rotating, blurring, and brightness/saturation/hue/contrast jittering, etc. For Task 2, they proposed a heuristic algorithm to predict the whole tumor area from the viable tumor area determined in Task 1 using a convex hull algorithm. We observed that even though the Jaccard index accuracy of the whole tumor area generated using the proposed heuristic method was low, the viable tumor burden value was reasonably good, ranking second in Task 2.

3.5. Ma et al. (4th place, team damo-AIC)

Ma et al. proposed another multi-scale ensemble scheme based on multi-task learning for predicting viable tumor masks and whole tumor masks. The proposed method used three levels (x200, x50, and x12.5) of WSIs to generate three outputs, which were then combined into one result using an ensemble method. For the highest resolution (200x) WSIs, a customized multi-task net-

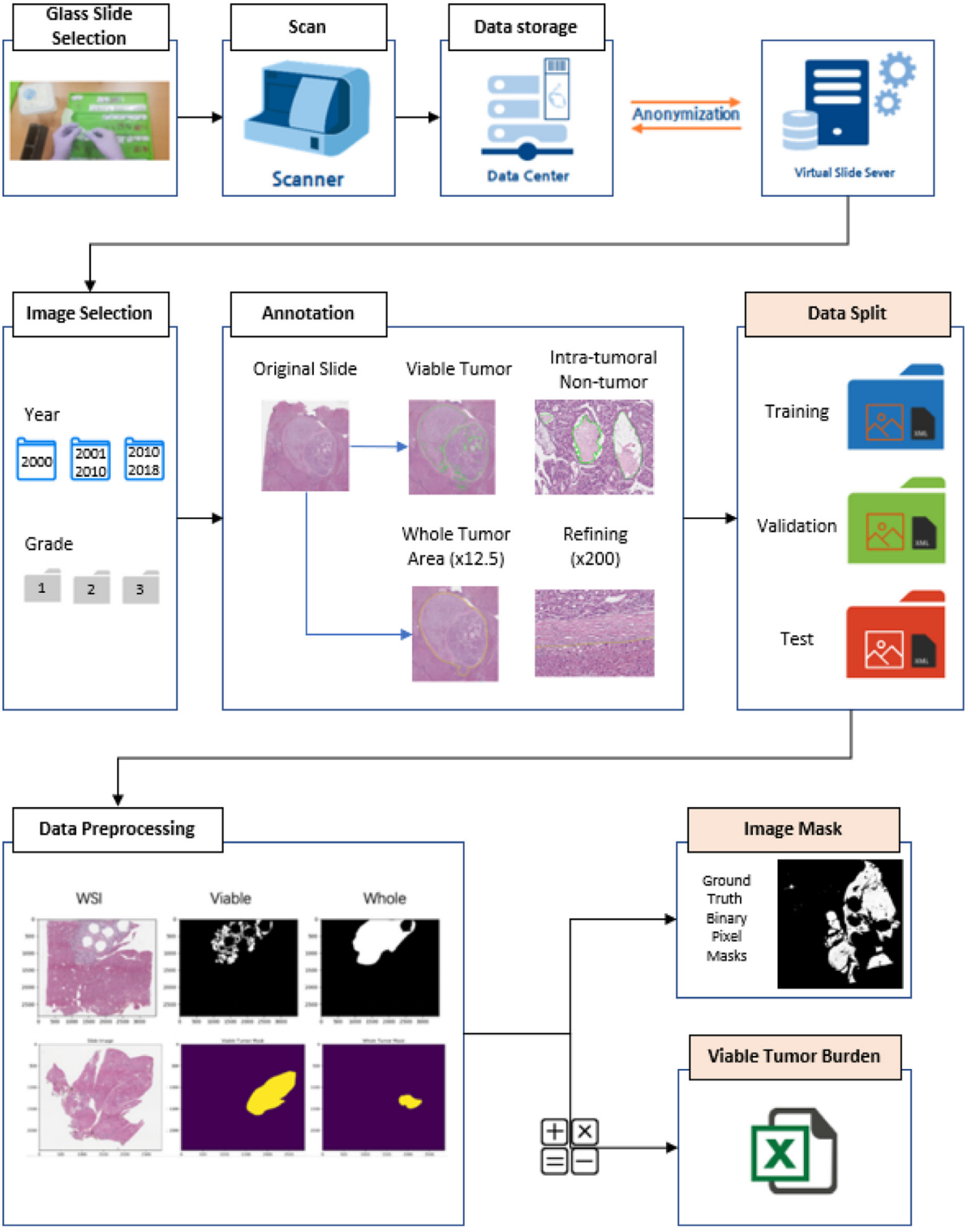

Fig. 2. The process of PAIP dataset preparation.

work including a ResNet101 (He et al. (2016)), backbone, and ASPP blocks (Chen et al. (2018)), was used, while a U-Net with a ResNet34 (He et al. (2016)) backbone was used for the two smaller resolution images. The cross-entropy loss was used for classification while the dice and focal loss (Lin et al. (2017)) were used for segmentation. They demonstrated that their proposed model outperforms the state-of-the-art models (e.g., U-Net and DeepLabV3+) using our PAIP challenge data.

3.6. Han et al. (5th place, team QUIIL)

Han et al. proposed the cascading of three deep neural networks to successively conduct patch-level liver cancer detection and pixel-level viable tumor segmentation. For liver cancer detection, the authors proposed a per-patch classification network based on DenseNet-121 (Huang et al. (2017)) wherein the input WSI is decomposed into 1024 × 1024 patches, and each patch is clas-

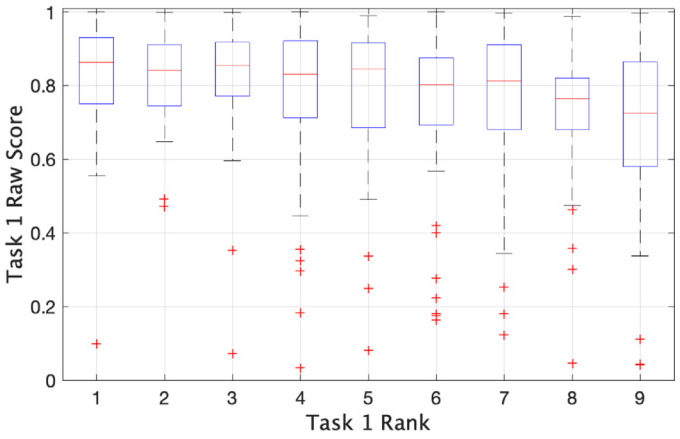

Fig. 3. The boxplot of the raw scores of Task 1 for the top nine teams.

sified as either tumorous or benign. For viable tumor segmentation, the patches classified as tumors in the previous stage were further classified into either viable or non-viable tumors using a classification network (involving the same architecture as shown above). Then, the viable tumor patches were further processed by the pixel-level segmentation network based on U-Net. For training, Adam optimizer (Kingma and Ba (2014)) and cross-entropy loss were applied to all networks. In addition, the focal-loss (Lin et al. (2017)) was adopted to account for the class imbalance in the data for the segmentation network. In Task 2, they estimated the viable tumor burden ratio using the results generated by the first and the third networks, which represent the whole tumor area (estimated per-patch) and the viable tumor area (estimated per-pixel).

4. Results

4.1. Results of task 1: Liver cancer segmentation

The final scores of Task 1 for the top nine teams were reported in Table 4 in the order in which they ranked. The top three teams achieved scores of approximately 0.75–0.79, which are substantially higher than 0.63–0.67 scores of the achieved by the fourth to eighth ranked teams. The p-value of the paired *t*-test between the top three versus fourth to sixth was 7.22e-3 and that of the independent *t*-test between the top three versus fourth to eighth was 9.78e-5, which means the difference is significant enough (see Fig. 6). The ninth team yielded a 0.53 score. Fig. 3 shows the boxplot of the raw Jaccard scores for the top nine teams without thresholding to show outliers. The top three teams have one to two outliers indicated by the red plus-sign markers, while the other lower-ranked teams have more than two outliers or a larger standard deviation. Almost all teams yielded low raw Jaccard scores that are lower than 0.2, which implies that the algorithms failed to detect viable tumors. However, the team 'Sen' did not yield such low raw scores for all test cases; and all cases for 'Sen' yielded Jaccard scores higher than 0.4.

Table 4
The final scores of Task 1: Liver Cancer Segmentation for the top nine teams.

Rank	Team	Score
1	newhyun00	0.7890
2	Sen	0.7772
3	MIRL-IITM	0.7503
4	Damo AIC	0.6718
5	QuIIL	0.6652
6	CUHK-Med	0.6625
7	DAISYlab@UKE	0.6596
8	COSYPath	0.6313
9	LRDE	0.5299

Fig. 4 illustrates the raw Jaccard scores of Task 1 for all individual test images, represented as boxplots for the top nine teams. This figure shows that there are some pathology images are easily segmented such as the test images with indices 4, 6, 8, and 9, among others, while there are also very challenging images to segment, such as the test images with indices 21, 24, 32, and 34, among others. Simple majority-voting based ensembles for the top three results (indicated by the red asterisks) and for the top nine segmentation results (indicated by the green circles) were reported. The ensemble of the segmentation masks of the top three teams resulted in improved overall raw scores from 0.82 (the best raw score of 'newhyun00') to 0.85 (top 1–3 ensemble). The extent of the outliers in terms of the lowest raw Jaccard score was improved, as the lowest Jaccard score of 0.47 obtained by team Sen was improved to 0.58. However, the ensemble of the segmentation

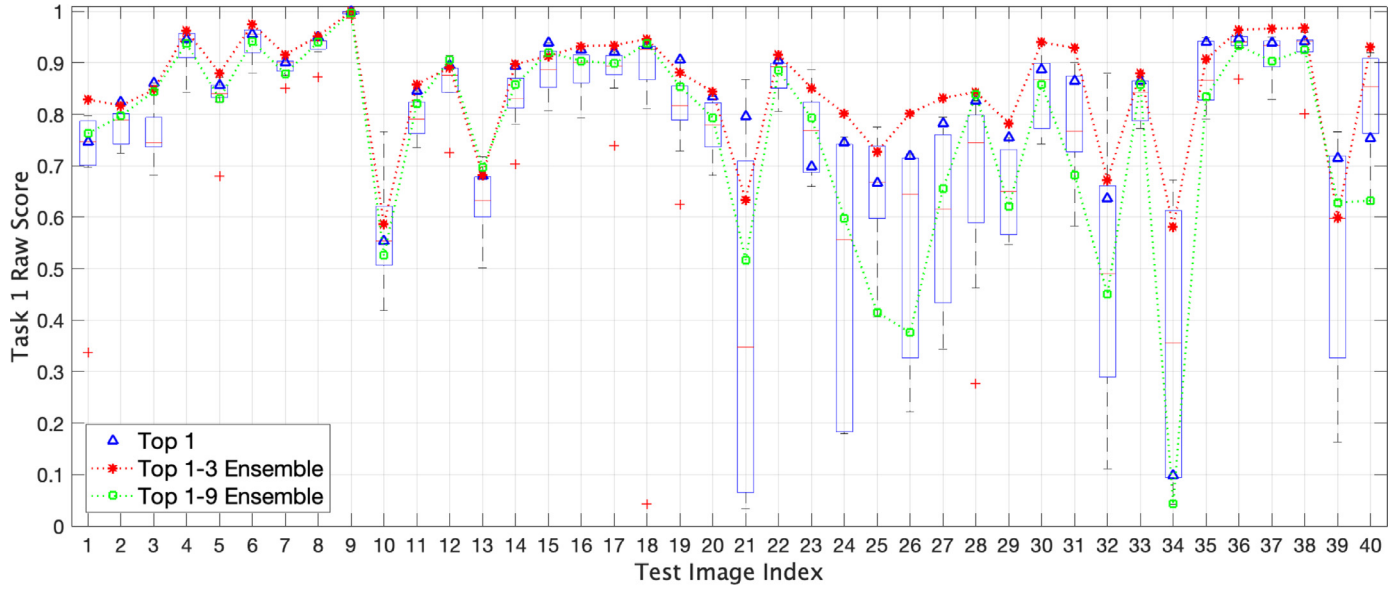

Fig. 4. The boxplot of the raw scores of Task 1 for all individual test images.

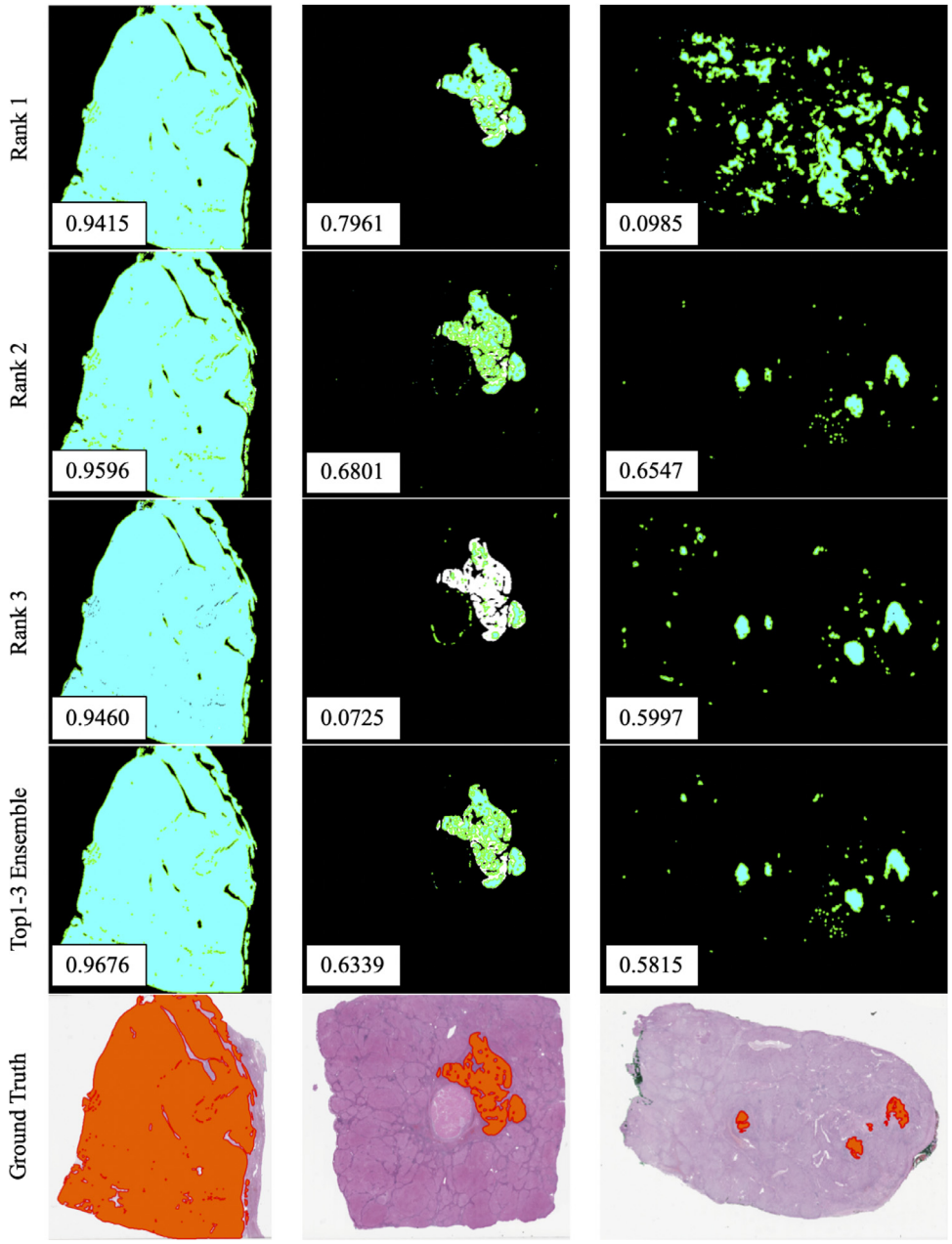

Fig. 5. Segmentation results for 3 selected test images for one high and two low average final scores whose test image indices are 38, 21, 34, respectively.

masks created using the results of all top nine teams yielded poor scores for many test images. Fig. 5 illustrates the three cases for one high- and two low- average final scores for the top three ranking teams (the first three rows) along with the top 1–3 ensemble results (the fourth row) with the ground-truth masks (white) and masks determined from the results (green). The last row shows the original pathology images with the ground-truth masks (red). In many cases, the top three methods yielded good segmentation masks, as shown in the first column of Fig. 5. However, there are often outlying results, such as the mask in the first row and the third column of Fig. 5 (indicating overdetection) and the mask in the third row and the second column of Fig. 5 (indicating under detection).

Table 5
The final scores of Task 2: Viable Tumor Burden Estimation for the top nine teams.

Rank	Team	Score
1	newhyun00	0.7528
2	MIRL-IITM	0.6337
3	QuIIL	0.6330
4	Damo AIC	0.6200
5	COSYPath	0.5969
6	CUHK-Med	0.5883
7	DAISYlab@UKE	0.5774
8	Sig-IPath	0.4624
9	blackbear	0.4335

4.2. Results of task 2: Viable tumor burden estimation

The final scores of Task 2 for the top nine teams are reported in the order of their ranks in Table 5. The top team achieved a score of 0.75, which is substantially higher than the 0.62–0.63 scores obtained by the groups ranked second through fourth. The groups ranked fifth through seventh achieved scores in the range

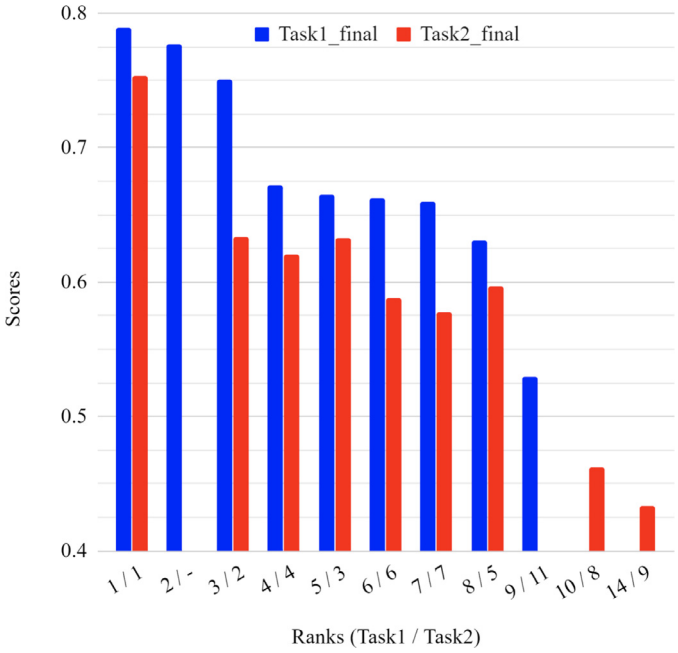

Fig. 6. Bar graph of the final top ranking scores.

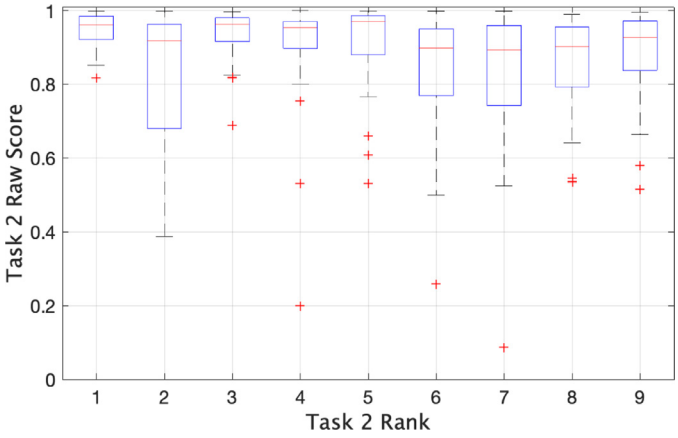

Fig. 7. The boxplot of the raw scores of Task 2 for the top 9 teams.

of 0.58–0.60, which are substantially higher than the 0.43–0.46 scores achieved by the groups ranked eighth and ninth. The p-value of one-way ANOVA between the first, second to fourth, fifth to seventh, and eighth to ninth groups (i.e., four groups) was 2.10e-5, which means the differences between the groups are significant enough (see Fig. 6).

Fig. 7 shows the boxplot of the raw scores for the top nine teams. The first ranked team achieved scores that are all higher than 0.8, while the third-ranked team achieved similar scores except for one outlier that is lower than 0.7. The second-ranked team yielded a wide range of scores from around 0.4 up to 1.0. Even though the raw scores of the third-ranked team were better than those of the second-ranked team, the final score of the second-ranked team was higher due to the weighting of the Task 1 results. The whole tumor estimation algorithm of the third-ranked team may possibly work better than the whole tumor estimation method of the second-ranked team. Fig. 8 shows the mean-normalized ratio, which is the estimated ratio minus the ground-truth ratio. Ideally, the mean-normalized ratio should be 0. The first-ranked team yielded good viable tumor burden estimation values for all cases thanks to the accurate viable segmentation results from Task 1. The third-ranked team also yielded good mean-normalized ratios similar to those of the first-ranked team.

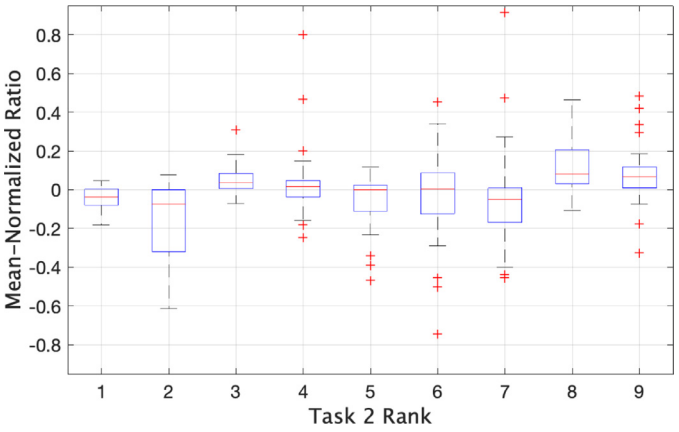

Fig. 8. The boxplot of the mean-normalized ratios of Task 2 for the top 9 teams.

Fig. 9 illustrates the raw mean-normalized ratios of Task 2 for all individual test images as boxplots. There are some pathology images for which the viable tumor burden estimation was relatively easy, such as the test images with indices 1, 7, 8, and 12, among others, while other images were more challenging, such as the test images with indices 21, 25, 28, and 39, among others. The simple majority-voting based ensembles for the top three ratio estimations (indicated by the red asterisks) and top nine (indicated by the green circles) ratio estimations are reported, respectively. The ensemble constructed using the estimations of the top three teams resulted in a slightly improved overall burden estimation from reducing the mean absolute error of 0.0530 (from team "newhyun00") to 0.0512, as well as reducing the number of outliers. However, the ensemble incorporating all of the top nine results yielded poor estimates for many test images and the mean absolute error was 0.0584. The ensemble results for Task 2 are therefore consistent with the ensemble results for Task 1.

Fig. 11 illustrates three cases for the first, third, and fourth ranking teams (the first three rows) along with the ensemble results for the top first, third, and fourth ranking teams (the fourth row) for the resulting masks (white indicates viable tumor, and green indicates whole tumors). The last row shows the original pathology images with the ground-truth masks. The results of the second-ranked team were not used in the creation of the ensemble results in this case because the raw score of this team was substantially lower than other teams in this ensemble. Additionally, the whole tumor segmentation results of the second-ranked team were severely down-sampled by 16 times, such that the creation of an accurate ensemble was not feasible. Note that this ensemble involving the top first, third, and fourth ranking teams was performed using the segmentation masks, rather than the final ratio estimations as shown in the fourth row of Fig. 11. In the first column, all results yielded good viable (shown in red) and whole (shown in purple) tumor masks, as well as a good viable tumor burden estimation that is close to 0.9548. The second and third columns of Fig. 11 illustrate the cases where relatively poor viable tumor burden estimations of the third- and fourth-ranked teams were compensated by majority voting-based ensembles of all three segmentation results to yield the ensemble estimates that are closer to the ground truths, respectively.

Lastly, the final scores between these two tasks are strongly correlated, as illustrated in Fig. 10, which shows an R^2 value of 0.873. It seems that our thresholding and Task 1 weighting allowed us to avoid obtaining high Task 2 scores with relatively poor segmen-

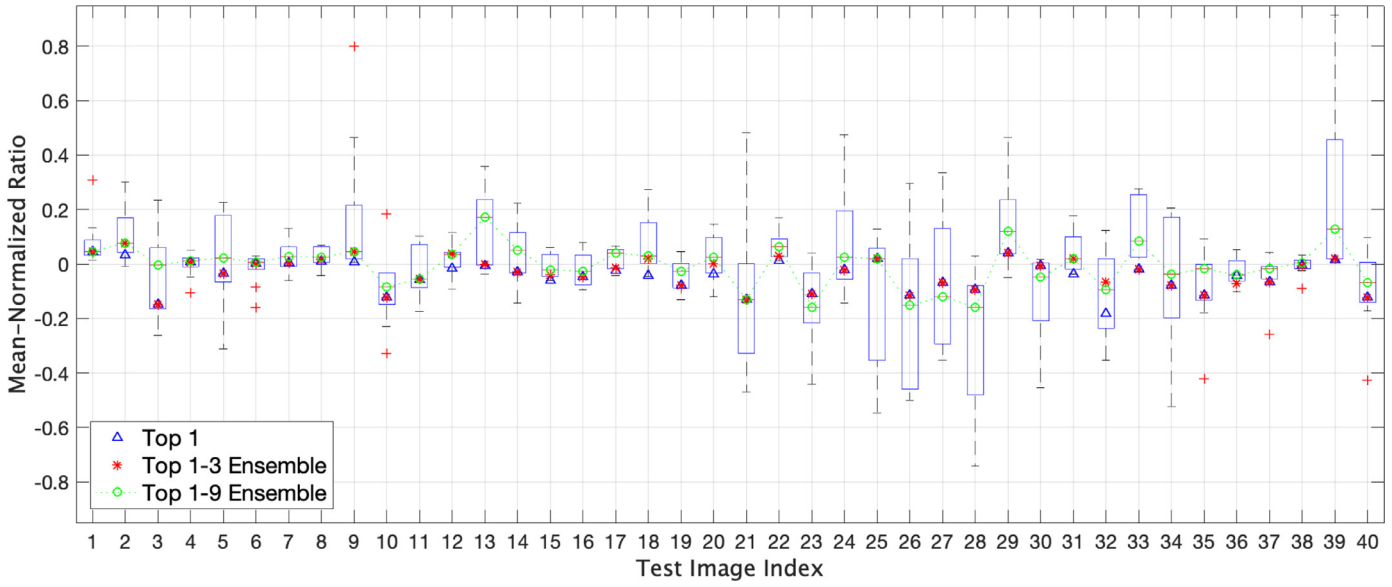

Fig. 9. The boxplot of the raw mean-normalized ratios of Task 2 for all individual test images.

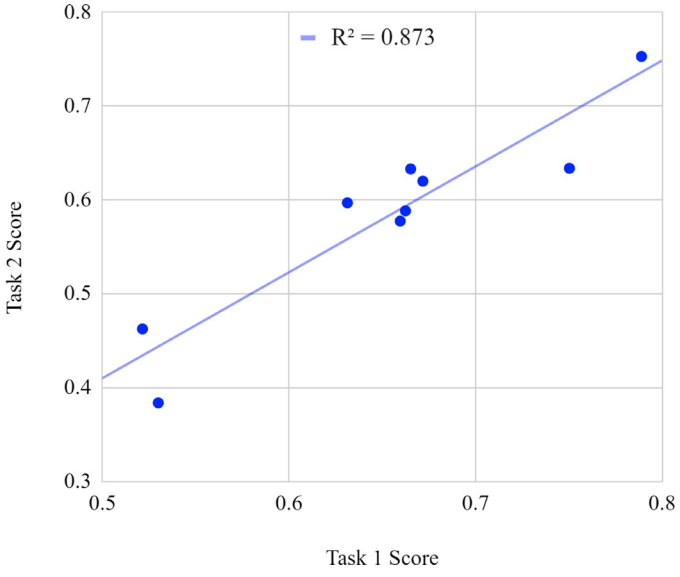

Fig. 10. Correlations between Task 1 / 2 final scores.

tation results for both viable and whole tumors, as illustrated in Fig. 11.

5. Discussion

5.1. Organization

There were about 794 registered users and over 231 data downloads in the PAIP 2019 Challenge. At the end of the contest, however, only 30 validation submissions and 28 test submissions were received. This drop in the submission rate is common in medical image challenges, which indicates a need to revise the design of future challenges to maintain the interest of the participants throughout the entire duration. The number of participants in Task 1 and Task 2 also differed. Task 1 had 27 participants submit their results, compared to only 17 submissions for Task 2.

The validation submission period allowed for the opportunity to identify technical problems that may occur during a relatively short test submission period. In the PAIP 2019 Challenge, in case the participants failed to upload their results due to technical problems during the submission of the test, we received an additional submission via email to be scored. Grand Challenge, the platform we used for this challenge, has limited upload capacity and this caused intermittent submission errors. The participants had to wait for 24h to determine whether or not the submission had been successfully uploaded, and then re-upload it if an error occurred.

5.2. Medical perspective

Anatomic pathology is a classical, yet an essential method for the confirmative diagnosis of a disease, especially neoplastic diseases. Digitalization of a pathological image makes image analysis possible and deep learning algorithms opened a new era of computational pathology because feature extraction based on cellular morphology would not be needed. Convolutional neural network algorithms using labeled images are superior to classical image analyses, which use limitedly assessed cellular features. Given the complex tissue images, it is practically impossible for a person to accurately annotate the object of classification. While preparing the data, there are a few things to consider. First is the level of annotation because the cellular level is most accurate but practically impossible to draw, and a realistic scale would be needed so that the algorithm accuracy is not affected. Second is the determination of the number of images is needed to generate the model. We used resected samples for training and had to consider which factors affect the modeling among the number of cases and the total patches of the image. The third is a matter of exact criteria and definition; for example, the tumor boundaries defining the whole tumor area do not have specific histological characteristics like the tumor cells; hence, it is not easy to distinguish them on glass slides, but they can be easily distinguished on gross examination. This is a matter of scale and some participants used different levels of images for modelling. The participants" results showed high accuracy at the case level, but all cases showed false-positive or false-negative results at the image level. For example, cases with low Jaccard scores were grade 1-HCCs, which were well-differentiated and similar in appearance to that of reactive hyperplasia of non-neoplastic hepatic nodules. (see Fig. 12) There was no difference in the grade distribution between the training and the test groups, but the higher the grade, the higher was the Jac-

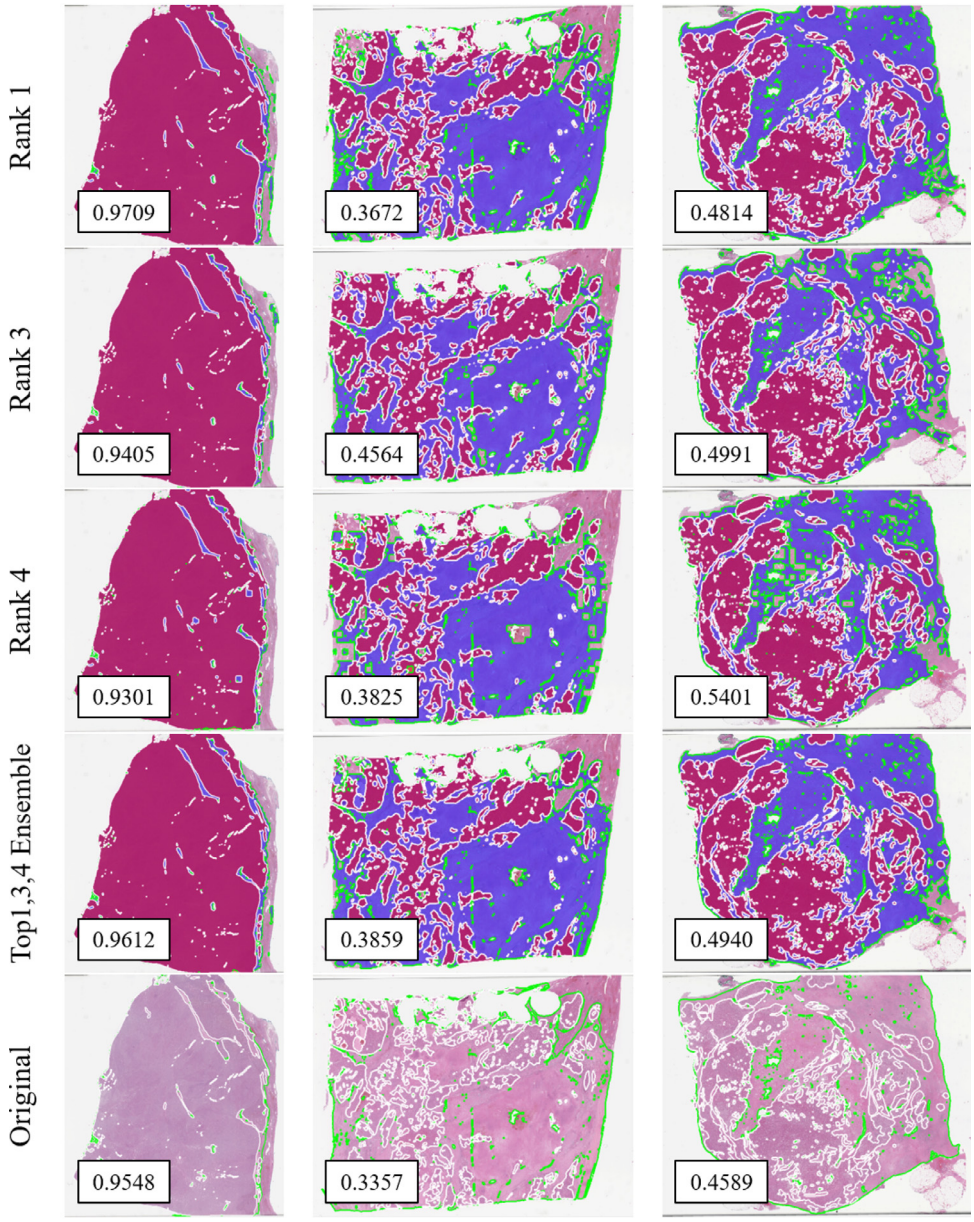

Fig. 11. Segmentation results for 3 selected test images whose test image indices are 38, 2, and 22, respectively.

Table 6
Statistics analysis results for each group of slides by cancer grades.

Features / Grade	Grade 1	Grade 2	Grade 3
mean ± SD	0.60 ± 0.11	0.73 ± 0.16	0.76 ± 0.18
Training [50]	7	23	20
Validation [10]	0	6	4
Testing [40]	3	15	22
Total	10	44	46

card score. (see Table. 6) The greater the difference was between the surrounding non-tumor tissue and the tumor, the higher was the accuracy of segmentation was. Given that well-differentiated tumors have a lower diagnostic agreement between the pathologists, a new approach is needed in this area.

5.3. Characteristics of the top ranked methods

We observed several common characteristics between the top-performing methods. One notable characteristic is the use of multi-scale inputs and ensemble methods (either explicitly with the outputs or implicitly with the network features). Many participants used two or three levels of WSI processing for the input data and combined the output results. In these methods, each input is processed by an independent network and the results are combined after analysis using either an averaging or a majority-voting ensemble method (newhyun00, MIRL-IITM, Damo-AIC). Some methods (DAISYlab@UKE and Sig-IPath) combined intermediate feature maps from the different networks to generate a single output at the end without an explicit ensemble operation. Another interesting characteristic is that some methods leveraged both multi-scale and multi-stage network features without multi-scale inputs. Team Sen proposed a combination of feature maps from different scales in the decoder network to improve the segmentation results. Jung et al. (newhyun00) employed a stochastic weight averaging method (Izmailov et al. (2018)) to average the network weights at different points in time during training to improve the convergence and accuracy of the results. We observed that about half of the top-ranked teams used a pre-trained network using the ImageNet

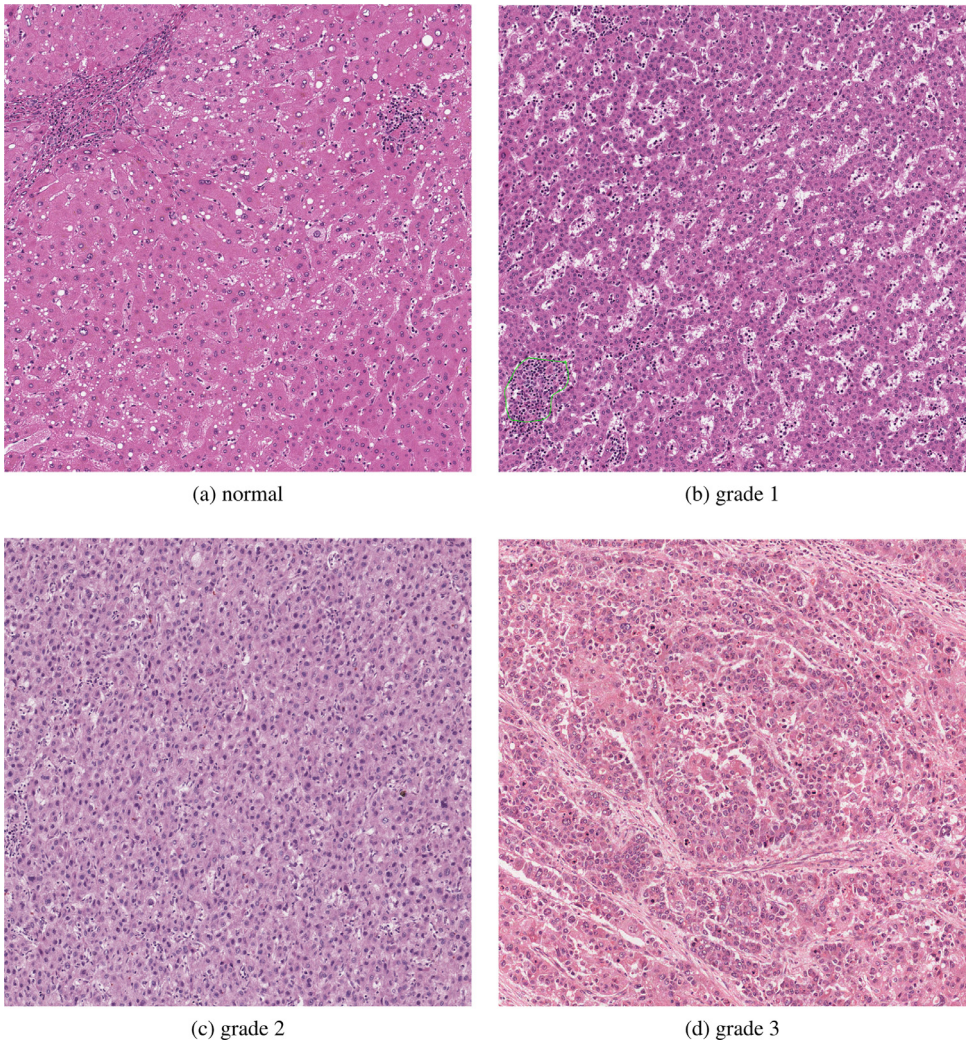

Fig. 12. Example patches from (a) normal, (b) grade 1, (c) grade 2, and (d) grade 3.

database, and most teams used various data augmentation methods, such as random geometric and color transformations.

5.4. Limitations

We created a large number of datasets containing liver cancer whole-slide images with annotations of viable tumors and whole tumor areas. Over the next two years, we plan to diversify this dataset by incorporating slides of different cancers with different annotations. We intend to invite more centers to participate in this project. We hope that the dataset and evaluation metrics we have provided will aid in the development and benchmarking of cancer diagnosis and segmentation. However, there are a number of problems that remain to be explored.

Firstly, patient confidentiality is a major barrier on the way to obtaining a valuable dataset. The PAIP database consists of anonymized data provided by deceased patients and patients who have signed consent and release forms. This is because sharing the whole-slide images that were obtained from patients that have not provided their consent violates privacy law, even after data anonymization. Therefore, the liver cancer data used in this study is limited and not representative of the larger range of liver cancer patients.

Secondly, the AI algorithms developed during the challenge have not yet been tested in terms of providing clinical diagnoses. These AI algorithms should be further evaluated and compared with the clinical diagnoses provided by doctors to determine their future applications. Considerable research is therefore required before these AI algorithms can be developed into software.

Thirdly, it will be difficult to establish ownership of the medical device developed from AI algorithms using PAIP's datasets. This creates a problem of ownership: which group owns patents to the medical device? Establishing ownership might prove difficult as equal claims can be made by the three stakeholders. Is it the intellectual property of the algorithms' developers, the hospital which provided the data set, or the manufacturer which developed the product?

Fourthly, the annotation of the whole-tumor boundary is not very accurate. As machine learning algorithms are known to rely on training data, the ground-truth should be accurately annotated (Aresta et al. (2019)). However, due to the complexity and high-cost of the cell-level annotation process, participants in this challenge had to use training data for the whole-tumor boundary area, which was not very accurate. Because there is no standard definition of the whole-tumor boundary area, it is difficult for pathologists to annotate the whole-tumor boundaries. As such, the aforementioned conclusions all require further refinement and correction in the light of future research.

Finally, the reproducibility of the results generated by the participants' algorithm was not high. We received the image masks or

the listed values as resulting predictions. Therefore, in addition to the methodology reported by the participants, we had to respond separately to the fact that various techniques could be applied to improve the quality of the results. After our challenge operation period, top competitors were asked to release their source code, but not all teams could release their code for various reasons. If the challenge platform is to be improved in the future, it is expected that the platform will generate and score the results by itself after receiving executable code to increase the reproducibility and evaluate the algorithm more transparently. If the platform can execute code, the shortage of hardware resources will no longer be a barrier to participation in the medical image challenges, which can also lead to increased accessibility to small budget participants.

6. Conclusion

The PAIP challenge was created to combat the lack of research that has been done to address liver cancer using digital pathology. Out of the 231 participants of the PAIP challenge datasets, a total of 64 were submitted from 28 team participants. There was a strong correlation between high performance on both tasks by teams, in which teams that performed well on Task 1 also performed well on Task 2. The submitted algorithms automatically predicted the segment of the liver cancer with WSIs to an accuracy of a score estimation of 0.78.

Three main conclusions can be drawn from this challenge. First, the participants' results were highly accurate at the case level, but all cases showed false positive and false negative results at the image level. Second, we found that the greater the difference between the surrounding non-tumor tissue and tumor, the higher the accuracy of segmentation. At last, the top-ranked teams achieved high performance with a pre-trained network using the ImageNet database. Various data augmentation methods such as random geometric and color transformation were commonly used among most teams.

The main takeaway points from the PAIP 2019 challenge are as follows:

1. To avoid participants overfitting their results, we recommend continuing with private test submissions. During the validation phase, participants would receive their results immediately. However, during the test phase, participants would be ranked before receiving their results.
2. We should extend the period of time for validation submission. This will allow for more participants to join the challenge and to increase participation in the final test submissions.
3. The final test ranking of the predictive results for medical images could be adapted to include various qualitative evaluations by medical experts. Using only a simple segmentation metric (e.g., the Jaccard score), clinically meaningful results might be marked with lower scores. Multiple experts participated in the qualitative evaluation.
4. The challenge rules on how to use external data are established. We must set standard guidelines for participants on whether to allow the use of external data according to the challenge. If allowed, there must be a clear rule on how to use external data to ensure all participants have access to the same resources.
5. Consider the reproducibility for analyzing the results. It is often difficult to reproduce the intermediate results of participants. If analytical data are required for subsequent studies, it is recommended to submit the code used in the challenge so that the algorithms can be analyzed in detail.

The challenge is designed to find the best possible algorithm to detect liver cancer in patients. We believe this will also further research into the applicability and trustworthiness of AI in clinical settings. We hope that this contribution will enable future quantitative studies of the progression and mechanism of the disease.

Declaration of Competing Interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

Jack Zeineh is a shareholder and management of PreciseDx.

CRediT authorship contribution statement

Yoo Jung Kim: Writing - original draft, Data curation, Investigation, Validation, Visualization, Conceptualization. **Hyungjoon Jang:** Writing - original draft, Data curation, Investigation, Validation, Visualization, Conceptualization. **Kyoungbun Lee:** Writing - original draft, Data curation, Investigation, Validation, Visualization, Conceptualization, Supervision, Project administration. **Seongkeun Park:** Writing - review & editing, Data curation, Investigation, Validation. **Sung-Gyu Min:** Data curation, Resources, Software, Validation. **Choyeon Hong:** Data curation, Resources, Software, Validation. **Jeong Hwan Park:** Data curation, Formal analysis, Validation. **Kanggeun Lee:** Data curation, Software, Validation, Visualization. **Jisoo Kim:** Data curation, Software, Validation, Visualization. **Wonjae Hong:** Data curation, Software, Validation, Visualization. **Hyun Jung:** Methodology, Resources. **Yanling Liu:** Methodology, Resources. **Haran Rajkumar:** Methodology, Resources. **Mahendra Khened:** Methodology, Resources. **Ganapathy Krishnamurthi:** Methodology, Resources. **Sen Yang:** Methodology, Resources. **Xiyue Wang:** Methodology, Resources. **Chang Hee Han:** Methodology, Resources. **Jin Tae Kwak:** Methodology, Resources. **Jianqiang Ma:** Methodology, Resources. **Zhe Tang:** Methodology, Resources. **Bahram Marami:** Methodology, Resources. **Jack Zeineh:** Methodology, Resources. **Zixu Zhao:** Methodology, Resources. **Pheng-Ann Heng:** Methodology, Resources. **Rüdiger Schmitz:** Methodology, Resources. **Frederic Madesta:** Methodology, Resources. **Thomas Rösch:** Methodology, Resources. **Rene Werner:** Methodology, Resources. **Jie Tian:** Methodology, Resources. **Elodie Puybareau:** Methodology, Resources. **Matteo Bovio:** Methodology, Resources. **Xiufeng Zhang:** Methodology, Resources. **Yifeng Zhu:** Methodology, Resources. **Se Young Chun:** Conceptualization, Writing - review & editing, Data curation, Visualization, Supervision. **Won-Ki Jeong:** Conceptualization, Writing - review & editing, Data curation, Supervision, Project administration. **Peom Park:** Conceptualization, Writing - review & editing, Data curation, Supervision, Project administration. **Jinwook Choi:** Conceptualization, Writing - review & editing, Supervision, Project administration, Funding acquisition.

Acknowledgments

This research was supported by a grant of the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea (grant number : HI18C0316)

Ethics approval: Approved by Seoul National University Hospital Institutional Review Board (IRB) (IRB No.H-1808-035-964).

The authors would also like to thank all the other PAIP 2019 Challenge participants: Pingjun Chen, Jack R. Collins (Team newhyun00), Quoc Dang Vu (Team QuIIL), Huaxin Yao (Team Damo AIC), Marcel Prastawa, Brandon Veremis, Nina Shpalensky and Gerardo Fernandez (Team COSYPath), Xi Wang (Team CUHK-Med), Maximilian Nielsen (Team DAISYlab@UKE), Yuxiang Ye and Yinan Chen (Team Sig-IPath), Guillaume Tochon (Team LRDE), Zhisheng Li, Li Wang and Zhongwei Zhang (Team blackbear), Haruki Kokubo

and Daisuke Komura (Team PRM-U), Woojoon Seok and Minsoo Yeo (Team U&I), Hatef Otroshi Shahreza, Ali Asghar Khani and Seyed Alireza Fatemi Jahromi (Team Phoenix), Tang-Kai Yin and Qi-Rui Fang (Team Yin&Fang), Gaoyuan Xu and Shuai Ye (Team Standing), Vo Thi Tuong Vi, ZhiWei Zhai, Adam Chudaś, Chao-Hui Huang, Haotian Cai, Jaehoon Jeong, Shen Yu-Jie, Hyeongsub Kim and Hongjun Yoon (Team DEEPNOID-POSTECH)

Appendix A. Source Code Availability

Table A1

Table A1
Finalists' source code availability.

Team	Source code availability	Link
newhyun00	√	https://github.com/newhyun00/paip2019_code
Sen	✗[a]	-
MIRL-IITM	√	https://github.com/koriavinash1/DigitalHistoPath
Damo AIC	✗	-
QuIIL	√	https://github.com/ChangHeeHAN/PAIP2019
CUHK-Med	✗[b]	-
DAISYlab@UKE	√[c]	https://github.com/RSchmitzHH/multiscale
COSYPath	✗[b]	-
LRDE	√	https://www.lrde.epita.fr/wiki/Challenges_codes
Sig-IPath	✗[b]	-
blackbear	✗[d]	-

[a] Pending: They will be participating in PAIP 2020 with the same network. They plan to open the source code after the competition.
[b] Not available due to intellectual property limits.
[c] It will be available after the team's manuscript publication process.
[d] Not available due to pending proposal for alternative project.

Appendix B. Related Links

- PAIP 2019 challenge website: https://paip2019.grand-challenge.org/
- PAIP 2019 forum: https://github.com/paip-2019/challenge
- PAIP website : http://www.wisepaip.org/

References

A.C., 2019. Residual cancer burden calculator. URL: http://www3.mdanderson.org/app/medcalc/index.cfm?pagename=jsconvert3.

Akbar, S., Peikari, M., Salama, S., Panah, A.Y., Nofech-Mozes, S., Martel, A.L., 2019. Automated and manual quantification of tumour cellularity in digital slides for tumour burden assessment. Sci Rep 9 (1), 1–11. doi:10.1038/s41598-019-50568-4.

Aresta, G., Araújo, T., Kwok, S., Chennamsetty, S.S., Safwan, M., Alex, V., Marami, B., Prastawa, M., Chan, M., Donovan, M., Fernandez, G., Zeineh, J., Kohl, M., Walz, C., Ludwig, F., Braunewell, S., Baust, M., Vu, Q.D., To, M.N.N., Kim, E., Kwak, J.T., Galal, S., Sanchez-Freire, V., Brancati, N., Frucci, M., Riccio, D., Wang, Y., Sun, L., Ma, K., Fang, J., Kone, I., Boulmane, L., Campilho, A., Eloy, C., Polónia, A., Aguiar, P., 2019. BACH: Grand challenge on breast cancer histology images. Med Image Anal 56, 122–139. doi:10.1016/j.media.2019.05.010.

Bejnordi, B.E., Litjens, G., Timofeeva, N., Otte-Höller, I., Homeyer, A., Karssemeijer, N., van der Laak, J.A., 2015. Stain specific standardization of whole-slide histopathological images. IEEE Trans Med Imaging 35 (2), 404–415.

Bhargava, R., Madabhushi, A., 2016. Emerging themes in image informatics and molecular analysis for digital pathology. Annu Rev Biomed Eng 18 (1), 387–412. doi:10.1146/annurev-bioeng-112415-114722.

Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H., 2018. Encoder-decoder with atrous separable convolution for semantic image segmentation. In: Proceedings of the European conference on computer vision (ECCV), pp. 801–818.

Cotoi, C.G., Khorsandi, S.E., Pleşea, I.E., Quaglia, A., 2012. Histological aspects of post-TACE hepatocellular carcinoma. Romanian Journal of Morphology and Embryology 53 (3 SUPPL.), 677–682.

fast.ai, 2019. URL: https://www.fast.ai.

Gurcan, M.N., Boucheron, L.E., Can, A., Madabhushi, A., Rajpoot, N.M., Yener, B., 2009. Histopathological image analysis: A Review. IEEE Rev Biomed Eng 2, 147–171. doi:10.1109/RBME.2009.2034865.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770–778.

Hu, J., Shen, L., Sun, G., 2018. Squeeze-and-excitation networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 7132–7141.

Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q., 2017. Densely connected convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 4700–4708.

Izmailov, P., Podoprikhin, D., Garipov, T., Vetrov, D., Wilson, A.G., 2018. Averaging weights leads to wider optima and better generalization. arXiv preprint arXiv:1803.05407.

Janowczyk, A., Madabhushi, A., 2016. Deep learning for digital pathology image analysis: a comprehensive tutorial with selected use cases. J Pathol Inform 7 (1). doi:10.4103/2153-3539.186902.

Kingma, D.P., Ba, J., 2014. Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980.

Li, X., Wang, W., Hu, X., Yang, J., 2019. Selective kernel networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 510–519.

Lin, T.-Y., Goyal, P., Girshick, R., He, K., Dollár, P., 2017. Focal loss for dense object detection. In: Proceedings of the IEEE international conference on computer vision, pp. 2980–2988.

Litjens, G., Kooi, T., Bejnordi, B.E., Setio, A.A.A., Ciompi, F., Ghafoorian, M., Van Der Laak, J.A., Van Ginneken, B., Sánchez, C.I., 2017. A survey on deep learning in medical image analysis. Med Image Anal 42, 60–88.

Liu, L., Jiang, H., He, P., Chen, W., Liu, X., Gao, J., Han, J., 2019. On the variance of the adaptive learning rate and beyond. arXiv preprint arXiv:1908.03265.

Longo, D.L., 2019. Augusto villanueva, M.D.,Ph.D. N Engl J Med 380, 1450–1462.

Holden, M., Smith, J., 2016. Preparing for the future of artificial intelligence. Technical Report. National Science and Technology Council.

Otsu, N., 1979. A threshold selection method from gray-level histograms. IEEE Trans Syst Man Cybern 9 (1), 62–66.

PAIP, 2019. Pathology AI Platform. URL: http://www.wisepaip.org.

Puybareau, E., Tochon, G., Chazalon, J., Fabrizio, J., 2018. Segmentation of gliomas and prediction of patient overall survival: a simple and fast procedure. In: International MICCAI Brainlesion Workshop. Springer, pp. 199–209.

Reinhard, E., Adhikhmin, M., Gooch, B., Shirley, P., 2001. Color transfer between images. IEEE Comput Graph Appl 21 (5), 34–41.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. Springer, pp. 234–241.

Sanjay Kakar, Chanjuan Shi, 2017. F. F. A. M. K. M. M.-K. P. J.-N. V. M. K. Washington. Protocol for the examination of specimens from patients with hepatocellular carcinoma.

Schmitz, R., Madesta, F., Nielsen, M., Werner, R., Rösch, T., 2019. Multi-scale fully convolutional neural networks for histopathology image segmentation: from nuclear aberrations to the global tissue architecture. arXiv preprint arXiv:1909.10726.

Seok, J.Y., Na, D.C., Woo, H.G., Roncalli, M., Kwon, S.M., Yoo, J.E., Ahn, E.Y., Kim, G.I., Choi, J.-S., Kim, Y.B., et al., 2012. A fibrous stromal component in hepatocellular carcinoma reveals a cholangiocarcinoma-like gene expression trait and epithelial-mesenchymal transition. Hepatology 55 (6), 1776–1786.

Song, J., Ge, Z., Yang, X., Luo, Q., Wang, C., You, H., Ge, T., Deng, Y., Lin, H., Cui, Y., et al., 2015. Hepatic stellate cells activated by acidic tumor microenvironment promote the metastasis of hepatocellular carcinoma via osteopontin. Cancer Lett 356 (2), 713–720.

Szegedy, C., Ioffe, S., Vanhoucke, V., Alemi, A.A., 2017. Inception-v4, inception-resnet and the impact of residual connections on learning. In: Thirty-First AAAI Conference on Artificial Intelligence.

Tan, M., Le, Q.V., 2019. Efficientnet: rethinking model scaling for convolutional neural networks. arXiv preprint arXiv:1905.11946.

Yao, F.Y., Kerlan, R.K., Hirose, R., Davern, T.J., Bass, N.M., Feng, S., Peters, M., Terrault, N., Freise, C.E., Ascher, N.L., Roberts, J.P., 2008. Excellent outcome following down-staging of hepatocellular carcinoma prior to liver transplantation: an intention-to-treat analysis. Hepatology 48 (3), 819–827. doi:10.1002/hep.22412.

Zhou, Z., Siddiquee, M.M.R., Tajbakhsh, N., Liang, J., 2018. Unet++: a nested U-net architecture for medical image segmentation. In: Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support. Springer, pp. 3–11.

MOIS, 2011. Personal Data Protection Laws in Korea.