This article has been accepted for publication in IEEE Journal of Biomedical and Health Informatics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2023.3327489

LYSTO: The Lymphocyte Assessment Hackathon and Benchmark Dataset

Yiping Jiao, Jeroen van der Laak, Shadi Albarqouni, Zhang Li, Tao Tan, Abhir Bhalerao, Shenghua Cheng, Jiabo Ma, Johnathan Pocock, Josien P.W. Pluim, Navid Alemi Koohbanani, Raja Muhammad Saad Bashir, Shan E Ahmed Raza, Sibo Liu, Simon Graham, Suzanne Wetstein, Syed Ali Khurram, Xiuli Liu, Nasir Rajpoot, Mitko Veta, Francesco Ciompi

***Abstract*— We introduce LYSTO, the Lymphocyte Assessment Hackathon, which was held in conjunction with the MICCAI 2019 Conference in Shenzhen (China). The competition required participants to automatically assess the number of lymphocytes, in particular T-cells, in images of colon, breast, and prostate cancer stained with CD3 and CD8 immunohistochemistry. Differently from other challenges setup in medical image analysis, LYSTO participants were solely given a few hours to address this problem. In this paper, we describe the goal and the multi-phase organization of the hackathon; we describe the proposed methods and the on-site results. Additionally, we present post-competition results where we show how the presented methods perform on an independent set of lung cancer slides, which was not part of the initial competition, as well as a comparison on lymphocyte assessment between presented methods and a panel of pathologists. We show that some of the participants were capable to achieve pathologist-level performance at lymphocyte assessment. After the hackathon, LYSTO was left as a lightweight plug-and-play benchmark dataset on grand-challenge website, together with an automatic evaluation platform. LYSTO has supported a number of research in lymphocyte assessment in oncology. LYSTO will be a long-lasting educational challenge for deep learning and digital pathology, it is available at https://lysto.grand-challenge.org/.**

***Index Terms*— Lymphocyte assessment, computational pathology, artificial intelligence, computer-aided diagnosis**

Manuscript received XXXXX; revised XXXXX; accepted XXXXX. This work was supported in part by European Union's Horizon 2020 research and innovation programme under grant agreement no. 825292 (ExaMode project, http://www.examode.eu), and from the Alpe dHuZes / Dutch Cancer Society Fund, grant number KUN 2014-7032. Corresponding author: Francesco Ciompi.

Yiping J. is with Nanjing University of Information Science & Technology, and was formerly with Department of Pathology, Radboud University Medical Center, Nijmegen (ping@nuist.edu.cn). Jeroen L., Francesco C. is with Department of Pathology, Radboud University Medical Center, Nijmegen, The Netherlands (francesco.ciompi @radboudumc.nl, Jeroen.vanderLaak@radboudumc.nl). Jeroen L. is also with Center for Medical Image Science and Visualization, Linköping University, Linköpi. Shadi A. is with Helmholtz AI, Helmholtz Zentrum München, 85764 Nuerherberg, Germany, and also with Faculty of Informatics, Technical University of Munich, 85748 Garching, Germany. Zhang L. is with College of Aerospace Science and Engineering, National University of Defense Technology, China, and also with Hunan Provincial Key Laboratory of Image Measurement and Vision Navigation, China. Tao T. is with Macao Polytechnic University, Macao, China. Abhir B., Johnathan P., Navid A. K., Raja M. S. B., Shan E. A. R, Simon G., Nasir R. is with Department of Computer Science, University of Warwick, Coventry, United Kingdom. Jiabo M., Sibo L., Shenghua Cheng, Xiuli Liu is with Huazhong Univ Sci & Technol, Wuhan Natl Lab Optoelect, Britton Chance Ctr Biomed Photon, China. Josien P.W. P., is with Medical Image Analysis Group, Department of Biomedical Engineering, Eindhoven University of Technology, Eindhoven, The Netherlands. Suzanne C. W. and Mitko V. is with Medical Image Analysis Group, Department of Biomedical Engineering, Eindhoven University of Technology, Eindhoven, The Netherlands. Syed A. K. is with School of Clinical Dentistry, University of Sheffield, Sheffield, United Kingdom.

Additionally, Jiamei Sun and Thomas Watson were participants of LYSTO and should be listed as co-authors, however we fail to contact with them to get signed author consent. Jiamei S. is with Information Systems Technology and Design Pillar, Singapore University of Technology and Design (SUTD), Singapore. Thomas W. is with GlaxoSmithKline Plc (GSK).

I. Introduction

Cancer and the host immune system have a complex, yet not fully understood, interplay. Over the years, clinicians and researchers in immuno-oncology have been investigating mechanisms involved in the tumor-immune microenvironment (TME), aiming at designing biomarkers that can capture a snapshot of such a scenario, and use those biomarkers to address one of the stringent questions in oncology: what to do next?

Over the years, the role of immune cells, and in particular the tumor-infiltrating lymphocytes (TILs), has increasingly been investigated[1]. Within the context of TILs in histopathology, two main research lines can be identified. The first line relies on the analysis of standard hematoxylin and eosin (H&E) stained histopathology slides and the quantification of a *TIL score*[2], estimated as the percentage of tumor-associated stroma region covered by lymphocytes and plasma cells. Several studies have shown that such a TIL score has prognostic and predictive value in breast cancer [3] as well as across a number of cancer types[4].

The second line relies on immunohistochemistry (IHC) to analyze T-cells, a subset of lymphocytes. Using IHC, specific types of cells can be identified in histopathology slides by targeting them via antigen-antibody interactions, and using a specific chromogen to distinguish them from other cells. In the context of lymphocyte assessment, the Immunoscore[5] was promoted to focus on T-cells that are positive to CD3 (all

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

This article has been accepted for publication in IEEE Journal of Biomedical and Health Informatics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2023.3327489

T-cells) and CD8 (cytotoxic T-cells) IHC markers, in particular at the tumor invasive front and in the tumor bulk.

Both the Immunoscore and TIL scoring approaches assess the density of immune cells as a biomarker, which therefore relies on the counting of lymphocytes. This task suffers from implicit variability and tediousness when performed by pathologists, suggesting the potential value of a computer-aided system. However, despite the simple nature of this task, it has been shown recently[6] that detecting lymphocytes in IHC goes beyond simply "counting dark-brown spots". Moreover, IHC slides in daily practice contain challenging regions such as dense clusters, possibly background staining, and presence of artifacts such as ink (see examples in Figure 1). Additionally, IHC also suffers from variation in tissue preparation, staining and scanning that is implicitly present across different pathology laboratories.

With the Lymphocyte Assessment Hackathon (LYSTO) as well as the benchmark dataset, we proposed and fostered the automated quantification of CD3 and CD8 positive cells in IHC images across different cancer types, including breast, colon, and prostate cancer. Hosted in 2019, this paper looks back at the organization, sample acquisition, and performance of developed frameworks during the event. We also reported recent progresses based on post-event submissions to our online platform.

Compared to previous challenges in this field, the LYSTO hackathon has two main novel aspects. First, it formulated the problem of cell counting in a weakly supervised learning fashion, where a single count is provided for each image, rather than exhausted annotations for individual cell. Second, it challenged participants to develop a solution in a short amount of time, namely a few hours, which justifies the 'hackathon' epithet, as well as its name, partly inspired by the word '*listo*', a Spanish term for 'clever', as well as 'ready/finished'.

Different from regular challenges in medical image analysis, LYSTO did not enforce specific restrictions on models, training schemes, or task types. As a one-day event in the form of hackathon or proof of concept, LYSTO encourages participants to focus on the problem, and try out any strategy that could be helpful. Some of the submitted methods have achieved on par performance with senior pathologists, suggesting the feasibility of applying automated methods for IHC evaluation. LYSTO will serve as a long-lasting educational dataset for machine learning and computational pathology.

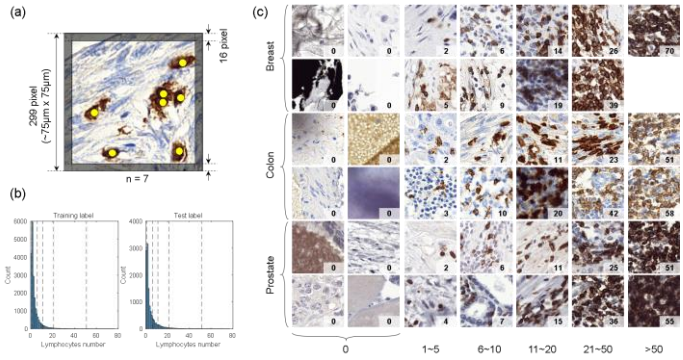

Fig. 1: The LYSTO dataset. (a) Example image patch used in the experiment. The label is calculated as the positive cells number in the central 267x267 pixel. (b) Label distribution in training set and test set. (c) Sample examples used in the hackathon. The number at the right bottom of each image indicates the reference standard.

II. Related Works

In this section, we discuss the evaluation of lymphocytes in H&E and immunohistochemistry (IHC) slides. Lymphocyte evaluation, as a subtask in IHC image analysis, may also be related to fine-grained hotspot detection[7] or multi-slide registration[8]. We will focus on IHC scoring and summarize challenge competitions or available datasets similar to LYSTO.

A. Lymphocyte assessment in H&E slides

Cell quantification via visual estimation is known to suffer from intra- and inter-observer variability. For this reason, recent studies proposed the use of deep learning to analyze digital pathology slides stained with H&E. In a recent work on breast cancer, high-TIL regions are recognized, and a deep learning model is then developed to quantify the TIL proportion [9]. The spatial arrangement of TILs has shown to be correlated with the tumor recurrence in lung cancer [10]. Although convenient, it is infeasible to recognize various lymphocyte subtypes in H&E slides, limiting more precise and quantitative analyses of the immune microenvironment.

B. IHC Scoring

Subtypes of lymphocytes can be identified through IHC staining. Since the recognition is mainly based on color, many methods can be generalized to other markers. We categorized current IHC scoring methods into color deconvolution-based methods and deep learning-based methods. The latter can also be further categorized into classification, segmentation, and detection frameworks.

Color Deconvolution-based Methods: The light absorbance contribution of Hematoxylin, Eosin, and Diaminobenzidine (DAB) can be separated in optical density space using Lambert-Beer's law [11]. The positive objects can be then recognized in the DAB channel, with different stain levels and further form an overall score [12]. Specific methods, for example, local adaptive threshold can be introduced for heavy staining cases, where the absorption is non-linear [13].

Except for thresholding, machine learning methods, such as shallow neural networks or decision trees can also be used for positive cell detection [14], [15]. For more complex shapes such as neurons, super-pixel segmentation is recommended as a preprocessing procedure[16]. For membrane staining patterns such as HER2, post-processing using image thinning was proposed in [17], which can accurately distinguish between HER2 0 and 1+. Color deconvolution can also be combined with deep networks used for cell detection or segmentation, as in the works on Ki-67 and HER2 scoring [18], [19].

Because of the simplicity in calculation, many open-source software is based on color deconvolution. ImmunoRatio was initially developed for scoring of ER, PR, and Ki-67 in breast cancer[20]. It based on color deconvolution, adaptive thresholding, and watershed segmentation. QuPath provides algorithms for multiple markers, such as CD3 and CD8, using peak detection after color deconvolution; it further construct a decision tree to determine p53 score [14].

Deep Learning-based Methods: Classification models are suitable for predicting image-level labels, and are therefore

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

This article has been accepted for publication in IEEE Journal of Biomedical and Health Informatics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2023.3327489

commonly used for classifying staining patterns of individual cells or estimating image-level IHC scores. [21] used Gamma mixture models to detect potential nuclei in Ki-67 images, and established a convolutional neural network (CNN) to classify single-nuclei image patches as positive or negative; the F1 score reaches 0.91 compared with pathologists. The positive class can be further extended to as weak, moderate, and strong classes, and forming image-level score using methods like voting. This strategy has been validated for HER2 scoring in breast cancer [22], [23], and simpler network (e.g., VGG or network of few layers) are recommended rather than complex networks such as Xception.

Similar to color deconvolution, semantic segmentation models produce pixel-level segmentation maps as intermediate results, which are further post-processed by thresholding or watershed to identify instances. This pipeline has been used in PathoNet to evaluate Ki-67 score in breast cancer [24]. Sometimes the positive area ratio can be directly used without instance segmentation, for example, evaluating the PD-L1 score in a previous work [25].

Detection models or instance segmentation models can directly obtain bounding boxes or instance masks, making it easy to count positive cells. YOLLO[26], based on the popular object detection model YOLO, is a detector for CD8+ lymphocytes and robust against brown artifacts. The locality sensitive model (LSM) introduces restrictions towards sparsity in nucleus center, which has been used for CD8+ lymphocyte detection[27]. [28] established a detector for CD3+ cells, tumor cells, and other cells based on RetinaNet, which was extensively validated in head and neck cancer, lung cancer, breast cancer, and gastric cancer. Nevertheless, challenges from stain artifacts, tissue folds, and dense regions are persisted.

Interestingly, some studies have found that combining semantic segmentation with traditional morphological algorithms can result in better performance than detection frameworks. In protein expression analysis of colorectal cancer, a U-Net-based segmentation model with watershed postprocessing outperformed Detectron2 pipeline [29]. Similarly, a U-Net with peak detection also outperformed YOLLO and LSM [27] in CD3 and CD8 scoring. Here, we emphasize the discovered trend, that more complex methods (including model structure and workflow) may not necessarily yield better results.

C. Related Datasets and Challenges

There are already some publicly available datasets for IHC image analysis, which include not only CD3, CD8 staining, but also Ki-67, PD-L1, and so on.

Gastric CD3: This dataset is released with a recent work for CD3+ lymphocytes infiltration in gastric cancer [30]. It contains 2717 patches, with a size of 70x70 pixels. The positive patches contain positive lymphocytes, while the negative patches represent the background or other cells.

SHIDC-B-Ki-67: The dataset consists of 2357 images with a size of 1228x1228 pixels, which are obtained from invasive ductal carcinoma of the breast, stained with Ki-67 and scanned at 400X [24]. The entire dataset contains annotations of 162,998 cell locations, including three categories, namely Ki-67 positive, Ki-67 negative, and lymphocytes.

Breast PD-L1: It is released with a ring-study of PD-L1 IHC of invasive breast cancer, which involves 109 IHC images scored by 31 pathologists[31]. Stain intensity scores are given on image-level, ranging from 1 to 4. The size of the images is 2160x2160 pixels, with a resolution of 0.524 μm/pixel.

HER2 Challenge: It is a challenge organized by the Tissue Image Analytics (TIA) Centre at Warwick University, which contains nearly 100 whole-slide images of HER2-stained breast cancer (100,000 x 80,000 pixels) [32]. Ground-truth labels are provided for each WSI, with scores of 0, 1+, 2+, and 3+.

Pan-cancer CD3: It provides 92 regions of interest from slides stained with CD3, each measuring 2mm^2, which involved head and neck squamous cell carcinoma, non-small cell lung cancer, triple-negative breast cancer, and gastric cancer[28]. The image resolution was 0.23μm/px, and the cellular annotation was performed jointly by pathologists and semi-automatic commercial software.

LYON: The LYON challenge (https://lyon19.grand-challenge.org/) [27] was proposed for a similar problem as LYSTO. LYON consists of 441 regions of interest (ROI). The aim of LYON is to provide an evaluation platform for comparison study, and neither training nor test labels are available. LYSTO data partially comes from LYON; in contrast, LYSTO samples are well-prepared, with specific reference standard. LYSTO can be seen as a twin challenge of LYON, with focus on cell counting.

Despite above IHC image datasets, there is still a lack of a benchmark dataset that can be easily utilized by researchers in the field of computer vision. This gap arises from various factors: firstly, cross-organ generality is preferred. Secondly, artifacts and cross-center data are extremely challenging, and not been fully considered yet. Finally, there is a lack of balance between generality and usability in terms of standardized image and tasks. Classification models developed based on single nuclei images must be applied on pre-detected nucleus, while fusion of local results introduce additional workflow. In summary, there is currently a lack of a standard IHC dataset that can be easily used, while possessing wide varieties in term of organs, centers, and patterns. The gaps drove us host the LYSTO hackathon.

III. LYSTO Hackathon

In this section, we describe the data, experiment design, computing infrastructure and evaluation platform in LYSTO.

A. LYSTO Datasets

We collected data from 83 whole-slide images (WSIs) of colon (n=28), breast (n=33) and prostate (n=22) cancer.

The images were collected and produced in a multi-centric fashion, including tissue preparation or staining from hospital in Eindhoven, Radboudumc, Rijnstate, Utrecht, Heerlen, AMC Amsterdam, as well as JBZ (Jeroen Bosch Ziekenhuis) and LabPON (Laboratorium Pathologie Oost-Nederland). All the slides were stained with either CD3 or CD8 immunohistochemistry, and most of the slides were stained in local institutes. Breast cancer cases involve 9 slides from LabPON and 3 slides from Radboudumc, while colorectal cancer cases involve 10 slides from Eindhoven, 5 slides from Utrecht, and 16 slides from Radboudumc, all corresponding to unique patients. For prostate cancer we included 11 patients

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

This article has been accepted for publication in IEEE Journal of Biomedical and Health Informatics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2023.3327489

from Rijnstate, with each patient having two slides, and they were split on patient-level during training and test division.

In order to introduce more variation in staining style, an additional set of triple-negative breast cancer (TNBC) cases was included [33], from which a subset of data consisting of 21 sections was generated and used in LYSTO. The 21 sections were collected from 3 TNBC patients (5, 5, 11 slides, respectively) and cut at the Radboudumc, therefore producing so-called "blank" (i.e., unstained) slides. Successively, these blank slides were sent for staining to 6 different pathology laboratories in the Netherlands, including Radboudumc, Rijnstate, Heerlen, AMC Amsterdam, JBZ, and LabPON. All the above slides were scanned using a Pannoramic 250 Flash II scanner (3DHistech, Hungary) at Radboudumc, and have a resolution of 0.24μm/pixel.

We split the slides into a training set (n=43 slides) and a test set (n=40 slides). Most of the slides are divided on patient-level. The high tissue homogeneity of 21 slides from three TNBC patients prevented us to do so. However, the multi-center staining nature in the TNBC subset allows us to eliminate histological factors and investigate the differences caused by staining difference among institutes. We will present the specific distribution of the data in the discussion section.

In order to effectively annotate the data, an expert initially selected an average of 11 regions of interest (ROIs) per slide, resulting in a total of 932 ROIs, with an average size of 2991 pixels by 4497 pixels (short side by long side), and an area of $1.33\text{mm}^2 \pm 3.15\text{mm}^2$ (mean ± std). ROI selection criteria were to not only include "regular" region of tumor epithelium and adjacent normal tissues, but also areas that were expected to be harder to analyze by deep learning algorithms, but anyway present in clinical diagnostic slides when considering the full slide, such as areas with densely distributed cells and artifacts (tissue folding, streaky blurring, brown staining on the background, blank slides, etc.).

Afterwards, three trained human analysts were asked to use ASAP software[34] to make point annotation for the center of each positive cell in the aforementioned ROIs. This resulted in the generation of over 170,000 annotations of cells[27]. The annotation process was carried out independently, but the analysts could discuss and decide together in difficult cases.

B. Sample Extraction

According to the ROIs and point annotations, we were able to generate patches and labels for training and validation. We used overlapped slide-window approach with a step size of 200 pixels to extract image patches of 299×299 pixels from each ROI. This corresponds to a tissue area of approximately 75×75μm. This specific patch size was selected to cover typical input shape of most popular CNNs, enabling using pre-trained models. In the context of a hackathon, where time is a vital factor, pre-training can reduce the convergence time.

The label of a patch y is defined as the number of annotated lymphocytes within it. Considering a point annotation could be present very close to the border, and the associated cell is only partly visible, we ignored annotations present within 4 μm thickness (approximately half the average size of a T-cell) at the border of a patch (Figure 1). For being compatible with classification task, we defined several bins for cell counts, namely 1~5, 6~10, 11~20, 21~50, 51~200, and >200. In the generation of the training set and test set, we tried to balance the labels according to these bins. We also collected plenty patches without the presence of lymphocytes, especially background stain region. To challenge participants with the dye artifacts in real-world applications, patches with $y=0$ were generated selectively according to the *brown score* proposed by [26]. The image source and label distribution of the LYSTO dataset is shown in Table 1.

TABLE I: Label distribution of LYSTO dataset

Properties		Training	Test
No. of slides	Breast	18	15
	Colon	13	15
	Prostate	12	10
Label value	Min	0	0
	Max	70	77
	Mean	3.11	3.92
No. of sample	0	4,208(21%)	2,915(24%)
	1~5	12,586(63%)	6,663(56%)
	6~10	2,008(10%)	1,260(11%)
	11~20	900(5%)	790(7%)
	21~50	290(1%)	323(3%)
	51~200	8(~0%)	49(~0%)
	>200	0(0%)	0(0%)
	Total	20,000	12,000

Given the slides in training set and test, we randomly selected n=20,000 and n=12,000 patches, respectively according to the rules above. In addition to the patch and corresponding label, the cancer type is also recorded as optional information in the training set.

C. External Validations

In parallel with collecting data above, we also collected a set of n=10 lung cancer slides from Radboudumc. Please note that the training set and test set of LYSTO do not contain lung images. Therefore, this external validation set can be used to assess the robustness and the generalizability to data from a different organ and different scanner. All these slides were stained with a CD8 marker and scanned with a Pannoramic 1000 scanner (3DHistech, Hungary), resulting in WSIs with a pixel size of 0.24 μm/pixel. Using the similar way as sample generation in LYSTO, we created n=54 ROIs and gathered annotations. The average physical size of these ROIs is 0.874 ± 0.641 mm^2 (mean ± std), and the ground-truth positive cell counts is 393 ± 412 (within entire ROI).

Additionally, we considered all the full LYON test set, this allows to test generalizability beyond single patches, especially when larger portions of challenging regions are present. Furthermore, running models on LYON allows comparison with expert pathologists using the observer study conducted in [27], where four pathologists were involved.

In order to perform validation on the external datasets, participants were asked to run their methods within a few months after the hackathon. We provide scripts that can processing patches in larger ROIs with slide-window fashion.

D. Timeline

The LYSTO experiment was a single-day event, held in conjunction with the Computational Pathology Workshop (COMPAY) at the MICCAI 2019 conference in Shenzhen

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

This article has been accepted for publication in IEEE Journal of Biomedical and Health Informatics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2023.3327489

(China). The hackathon was organized based on three main steps.

First, approximately one month before the event, a small dataset of *n*=4,000 labeled patches were released publicly via LYSTO website. The aim of this *pilot* dataset was to let potential participants get familiar with the data format that will be used during the event, and start coding pipelines that could be reused and modified during the event.

Second, the final official training set containing *n*=20,000 patches was released via the hackathon website three days before the event for the convenience of downloading.

Finally, the formal test set containing *n*=12,000 patches was solely released *on-site* via external storage units, and were manually distributed to participants. After the event, both training and test set were released publicly via the Zenodo platform. (https://zenodo.org/record/3513571)

E. Rules

As an application-driven challenge, LYSTO encourages participants to try out various solutions; therefore, no specific restrictions were enforced on model architecture, training schemes, data usage, or computing resources. We reformulate the problem of cell counting as a classification problem using pre-defined bins. This means that participants can solve the problem by using either classification, regression, or detection frameworks. Meanwhile, no restrictions were imposed regarding the data. Participants were allowed to reuse any materials in the community or append their in-house annotations. In summary, LYSTO is an open challenge, in which one can explore the most effective direction for future investigation towards cell counting task in IHC image.

F. Performance Metrics

In order to make the metric compatible with classification, regression, and detection frameworks, we make LYSTO as a patch classification problem using the 7 types of bins defined above (from 0 to >200 positive cells). In practical, pathologists will not identify and count individual cells in whole-slide images. To measure the consistency with reference standards, and penalize distinct errors (e.g., predict a patch with 50+ positive cells as none), we use quadratic weighted kappa (QWK) coefficient as the main performance metric on the LYSTO test set and external lung validation set. Meanwhile, QWK may also mitigate harmless error caused by observer variability.

Moreover, since the intervals defined in LYSTO is same as that in LYON, we are able to compare our results with the reader study described in LYON[27]. For this purpose, we report sensitivity in the LYON test set.

G. Baseline Results

We provided a baseline prior to the event, and submissions were thought to be valid if only it outperforms the baseline. The baseline was built with a decision tree using MATLAB (The MathWorks Inc., MA). Specifically, we extracted DAB channel of a patch, and use statistics including maximum, minimum, mean value, standard deviation, and percentiles of intensity as features. The features were then used to build a classification and regression tree (CART) to predict patch label. Using different prune levels, the test set performance ranges from about 0.628 to 0.649. In the end, a prune level of 1800 was used, which got 0.635 test set QWK (Figure 2). The baseline result and description were made available via website. In the end, all the on-site participants exceeded this baseline.

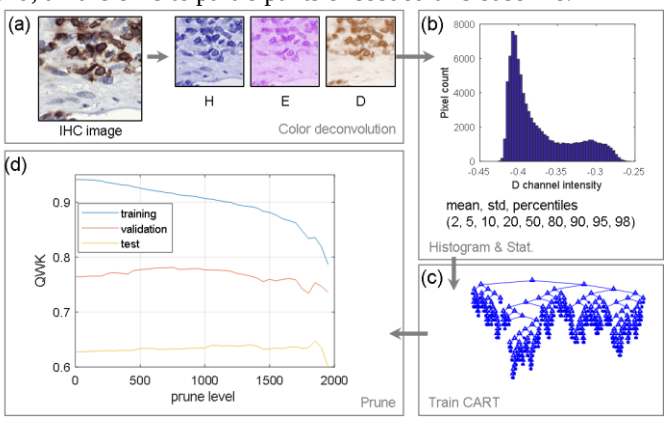

Fig. 2: The LYSTO baseline. (a) Color deconvolution; (b) Patch-level DAB channels statistics; (c) CART built with MATLAB; (d) Model performance and hyperparameter tuning.

H. Computing Resources

During the on-site event, we provided participants access to a dedicated GPU and storage on a cloud-based NVIDIA DGX-1 device, sponsored by NVIDIA. The official training set and test set were pre-loaded to the storage of the DGX-1. Additionally, participants were allowed to use local resources (e.g., their own laptop) as well as remote resources without any restriction.

I. Evaluation Platform

We implemented an automatic evaluation procedure, and released it via https://lysto.grand-challenge.org/. Participants were required to submit predictions with a single CSV file, and QWK scores can be calculated automatically. Examples of submission format were also provided upfront.

IV. Methods

During LYSTO, participants were permitted to form teams. Ultimately, five teams attended the on-site event and submitted their algorithms. After the event, we requested team leaders to provide a brief description of their methods. In this section, we outline the primary components of the developed methods, including 1) pre-processing, 2) data partitioning, 3) model architecture, and 4) training strategies.

A. Team 1 (GSK)

Preprocessing: Use the center 267×267 pixels of raw image as inputs. Patches were then normalized using ImageNet statistics. Horizontal and vertical flip, contrast and brightness adjustment (±20%, with p = 0.75) were used for augmentation.

Data split: The 20,000 training patches were stratified by the bins, and were further split into training and validation set with 3:1 randomly using built-in scikit-learn method.

Model architecture: The model is a multi-task network with a pre-trained ResNet-50 backbone. The last two layers of the ResNet were replaced with 4 convolutional layers (channel=2048, 1024, 512, 256, respectively, kernel size =3, padding =2, dilation=2). Afterward, regression and classification task branches were added. The two branches are both consisted by an adaptive pooling layer, a flatten layer, a Batchnorm1D layer, a dropout layer (*p*=0.25), ReLU activation,

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

This article has been accepted for publication in IEEE Journal of Biomedical and Health Informatics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2023.3327489

a linear layer (64 neurons), a Batchnorm1d layer, a dropout layer (p=0.25), and the final linear layer (with 1 or 7 neurons, for regression and classification task, respectively).

Training: Adam optimizer and one-cycle learning rate scheduler[35] implemented in fast.ai was used. The backbone was frozen in the first 20 epochs, using a maximum learning rate of 1e-3. Afterwards, the entire network became trainable for an additional 25 epochs. In the later phase, the maximum learning rate of backbone and task-specific layers is 1e-6 and 1e-4, respectively. The QWK scores calculated from the classification branch were monitored, and the regression branch is used for the final prediction.

B. Team 2 (HUST)

Preprocessing: The center 267×267 pixel region of the original image was used, and augmented by vertical and horizontal flipping, rotating in $n \times 90°$, and perturbating the brightness within a small range of values.

Data split: The 20,000 training patches were divided into ten folds at random, with an original intention to pick the best model using ten-fold cross-validation. However, during the hackathon it was decided to average the outputs of the 10 models trained during the cross-validation.

Model architecture: The model is a tailored ResNet-101 network for regression task. The last layer of the ResNet is removed, and a series of layers, including a global max pooling layer, a fully-connected layer (64 neurons) with ReLU activation, and a single-channel output layer, are attached.

Training: The model is trained with regression task, using mean squared error as loss function. The model was optimized with Adam optimizer (learning rate = 1e-3, step decay of 0.1 every 1500 iterations; exponential decay rates of 0.9 and 0.999 for the two moment) for a total of 6000 iterations with a batch size of 64.

C. Team 3 (TIA Warwick)

Preprocessing: The images are firstly reflected padded to the shape of 302×302. The pixel intensity is normalized within the range [0,1]. During training, flipping, contrast, brightness, median blur, Gaussian blur, and Gaussian noise are used for data augmentation.

Data split: Our method contains two networks, trained with segmentation and regression tasks, respectively. For the segmentation pretraining, we fully annotated the pilot training dataset using the ASAP software. In each patch, we ensured that there is an agreement between the number of annotated lymphocytes and the reference standard count. This dataset was split by the ratio of 7:3 into training and validation sets. Part of the segmentation network was reused in the final regression network, which was trained with the on-site 20,000 training images using five-fold cross-validation.

Model architecture and Training: Initially, a HoVer-Net [36] model was trained to perform instance segmentation of positively lymphocytes. The model was trained in two stages. In the first stage, the ResNet-50 encoder was initialized with weights pre-trained on ImageNet, and only the decoders were trained. In the second stage, both the encoder and decoder were trainable. The segmentation model was trained using Adam optimizer with an initial learning rate of 1e-3 and a batch size of 8 on each GPU.

After the training for segmentation task, the decoders are removed, and a series of 3×3 convolution, a max-pooling layers, an additional global average pooling layer, and a 1×1 convolution layer are added. The network output is a single value that regress the number of positively cells. In other words, the HoVer-Net encoder is used as a pre-trained network. The regression network is trained using Adam optimizer with an initial learning rate of 1e-3 and a batch size of 8 on each GPU. Mean absolute error is used as the loss function. The final prediction is the average of five cross-validation models.

D. Team 4 (TU/e)

Preprocessing: The input is mostly original patch size of 299 ×299 pixels, and zero-padded to size of 331×331 pixels if necessary. Images are augmented by random translation, rotation, scaling, shearing, flipping, and color channel shifting

Data split: The data split was done at the WSI level, ensuring that the validation and test set contain at least 1 WSI from each of the three tissue types. Out of the 43 unique WSIs, 32 were used for training, 5 for validation, and 6 as a test set.

Model architecture: Models with NASNet, Inception-ResNet-v2, Xception, SE-Net-154 and SE-ResNeXt-50 backbones were tested individually.

Training: All the backbones were pre-trained using ImageNet, and optimized with momentum SGD optimizer (learning rate 0.01, momentum 0.9, cosine annealing decay) for 50 epochs. The batch size ranges from 8 to 18, depending on GPU memory. Prediction is obtained by taking the median of the predictions of all single models.

E. Team 5 (mi2rl)

Preprocessing: Images are first split into two subsets by DAB channel, and the ones with fewer DAB are stain normalized. The center 267×267 pixels are used as input, with random rotation, flipping augmentation. Being aware that no sample with more than 200 positive cells are given, we take patches with high DAB response to generate new samples for that category.

Data split: 16305 image patches were used for training, and the rest was used for validation. The two sets are independent on slide-level. Instead of using raw bins, we cluster patches with their label to get more bins, ensuring that each bin has more than 50 samples. Image samples within the same bin were with the same label and the mean of the lymphocyte numbers per bin was used as the prediction.

Model architecture: The model is a classification model with DenseNet121 backbone, and attached classification layers. The raw and normalized images lead to two feature sets obtained from the end of backbone. The features are concatenated, and fed into another fully-connected layer for classification.

Training: The model was trained using AdamW optimizer (learning rate 1e-5, weight decay 0.05) for 10200 iterations with batch size 64. The loss was the distance between the median of the predicted and reference bins.

V. Results

A. "On-site" results

The on-site results of the five methods are reported in Table II in terms of QWK. In Figure 3, we depict the scatter plots of

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

This article has been accepted for publication in IEEE Journal of Biomedical and Health Informatics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2023.3327489

predictions versus ground-truth (a), the Sankey diagram of predictions (b), and examples of patches misclassified by all teams (c), grouped per bin in each row. From the QWK values, we see that most methods achieved comparable performance, with QWK>0.922, except for the mi2rl method (QWK=0.824).

TABLE II: Leaderboard of LYSTO event

Team name	On-site	Rank	External (lung dataset)	Rank
GSK	0.9270	1	0.9680	2
HUST	0.9247	2	0.8595	4
TIA Warwick	0.9229	3	0.9798	1
TU/e	0.9224	4	0.9652	3
mi2rl	0.8241	5	0.4678	5
Baseline	0.6350	-	0.8579	-

According to the Sankey diagram, the samples with label '0' and '1~5' are relatively easy, as majority samples are correctly predicted by all the teams. Such easy samples take 68.5% of the test set. In contrast, 4.0% of the samples are misclassified by all the teams, with some examples given in Figure 3(c). The misclassification is often correlated with background staining, resulting in strong DAB signal with few or none positive cells. Another typical case is partial membrane staining, which can be recognized by pathologists, but missed by most methods. The presence of artifacts (e.g., out-of-focus or ink), cluster of cells also lead to difficulties. More specifically, automated methods prone to underestimate cell number, especially when ground-truth count grows (Figure 3 (a)).

B. Post-event Submissions

The LYSTO hackathon remains open for new submission after MICCAI 2019. By now LYSTO has 667 registered users and receives 399 valid submissions. The highest QWK metric reaches up to 0.9331, which is higher than the top on-site group (GSK, 0.9270). According to gathered descriptions, newly submitted methods acquire similar techniques to the on-site groups. Participants are likely to use ResNet-18, ResNet-50, ResNeXT or U-Net as backbone. The task of classification, regression, or a combination of both are mostly used, with cross-entropy loss, and mean square error or Huber loss.

C. External Validations

After the on-site event, we asked the five teams to apply their methods on the two external validation datasets. In order to evaluate methods on ROI, local images generated by 16-pixel overlapped slide-window were evaluated, and the count

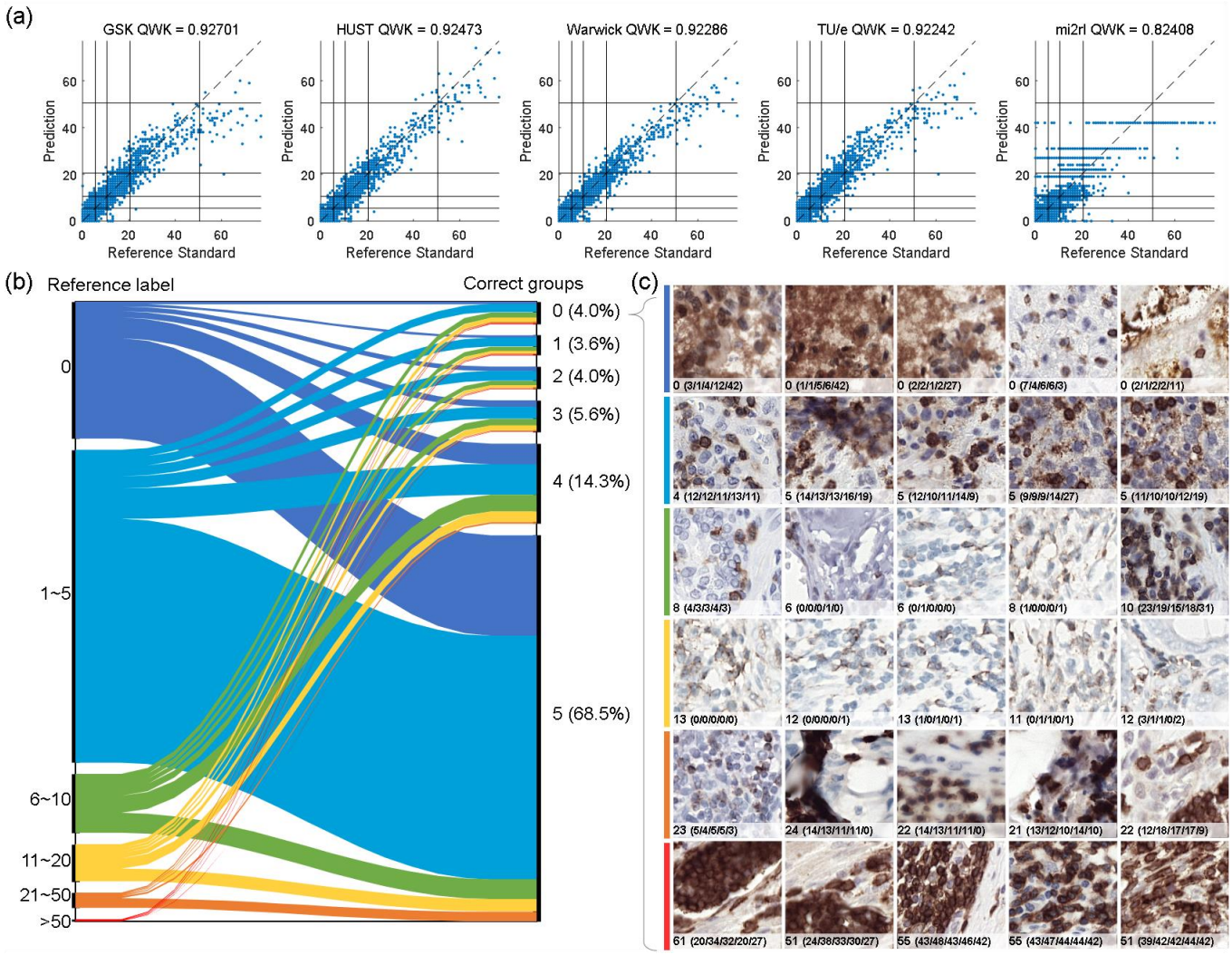

Fig. 3: On-site results. (a) Scatter plots of prediction versus reference standard. (b) Sankey diagram ground-truth versus number of groups that got correct prediction. (c) Examples of those 4% samples misclassified by all five groups.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

This article has been accepted for publication in IEEE Journal of Biomedical and Health Informatics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2023.3327489

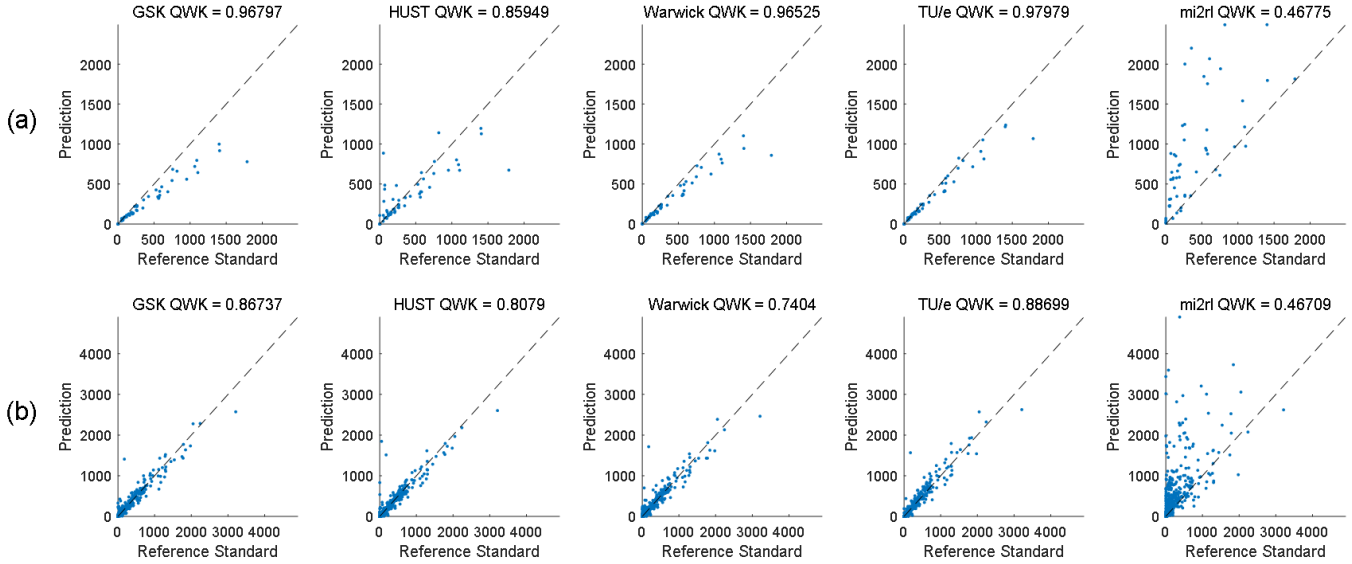

Fig. 4: Results of external validations. (a) The lung cancer cohort from Radboudumc. (b) The LYON test set.

numbers are summed up as ROI-level prediction. With larger field of view, these validations focus on model performance when applied to more representative of tissue morphology in whole-slide images used in routine clinical practice.

Lung dataset: The QWK values of the lung dataset are reported in Table II. The overall performance trend is similar to that in the LYSTO test set, with most methods achieving QWK>0.9. After ranking with average QWK (on-site and lung dataset)), GSK remains the best, followed by TIA Warwick, which achieved the best performance on the lung dataset.

TABLE III: Comparison with reader study in LYON

Label	0	1 ~5	6 ~10	11 ~20	21 ~50	51 ~200	>200	All
P1	0.78	0.11	0.25	0.15	0.32	0.71	0.54	0.41
P2	0.96	0.17	0.20	0.15	0.27	0.58	0.73	0.44
P3	0.78	0.28	0.15	0.20	0.32	0.48	0.35	0.37
P4	0.96	0.33	0.25	0.15	0.55	0.65	0.43	0.47
Average	0.87	0.22	0.21	0.16	0.37	0.60	0.51	0.42
[27]	0.30	0.44	0.30	0.35	0.54	0.76	0.92	0.52
TU/e	0.36	0.67	0.32	0.19	0.58	0.81	0.95	0.55
HUST	0.24	0.50	0.23	0.43	0.55	0.81	0.93	0.53
GSK	0.08	0.44	0.23	0.24	0.63	0.81	0.94	0.48
TIA Warwick	0.04	0.28	0.23	0.14	0.52	0.78	0.95	0.42
mi2rl	0.16	0.06	0.00	0.05	0.22	0.42	1.00	0.27

LYON Dataset: The performance on LYON can be seen as a generalization of patch-level performance of LYSTO, with more artifacts and cell clusters. In consistence with the reader study in previous study [27], the performance here was measured by sensitivity rather than QWK. The results are summarized in Figure 4 and Table III, where P1 to P4 stands for four pathologists involved in the study. In this dataset, TU/e got the best sensitivity. Notably, TU/e and HUST both got higher sensitivity than the method presented in [27], using fully-supervised point annotation. Moreover, the best four groups got better or comparable performances than the average of pathologists. These facts imply the feasibility to develop a human-level model using weakly-supervised patch-level label.

Moreover, algorithms and humans seem behave differently. Particularly, automatic methods prone to achieve higher sensitivity than pathologists when the number of lymphocytes grows. In contrast, human can easily distinguish positive cells from background staining or artifacts, which are usually challenging for computer algorithms.

VI. Discussion

LYSTO has witnessed the success of deep learning in medical image analysis, with participants extensively utilizing deep models such as ResNet, SENet, and DenseNet. By providing standardized data, LYSTO enables researchers to focus on specific problems and spend less time on coding, interface, and debugging. LYSTO gives a typical example demonstrating that standardized data can lead to clinically applicable solutions in a very short time.

Another benefit of standardized data is to allow participants time to explore a wide range of strategies or solutions. For instance, GSK adopted a multi-task learning strategy, using classification and regression tasks simultaneously. TIA WARWICK supplement extra manual annotations on a small part of the training set for pre-training. Mi2rl adopted stain standardization method, and defined more bins. TU/e used models with different architectures for ensembling, which often leads to performance gain[37]. Both GSK and TIA WARWICK adopted a two-stage training strategy[38], [39], where top layers or task layers are trained first with backbone frozen, and the entire network are trained together later. Limited by time duration, participants were unbale to finish ablation studies. Nevertheless, these techniques have already been widely employed in this field. As complete solutions, these methods have achieved comparable accuracy to human experts.

Despite the wide differences in the submitted methods, there are still noteworthy commonalities between them. The pair-wised prediction is shown in Figure 5. The Pearson correlation coefficients among the first four groups are all above 0.95. The ml2rl group used discrete prediction, which is different from others, resulting in relative lower Pearson coefficients to others, ranging from 0.86 to 0.88. Interestingly, all these values are higher than the correlation compared with ground-truth (0.84), indicating the commonalities between

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

This article has been accepted for publication in IEEE Journal of Biomedical and Health Informatics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2023.3327489

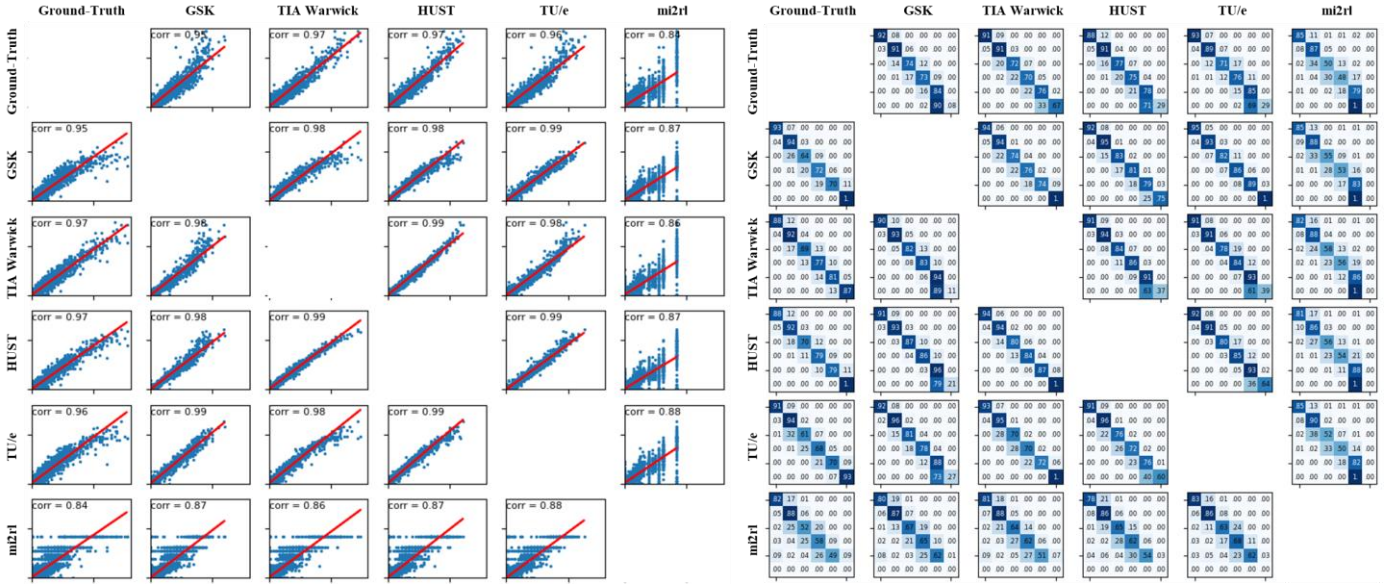

Fig. 5: Pairwise prediction correlation of on-site submissions, gives in count (scatter plot on the left) and bins (normalized recall matrix on the right).

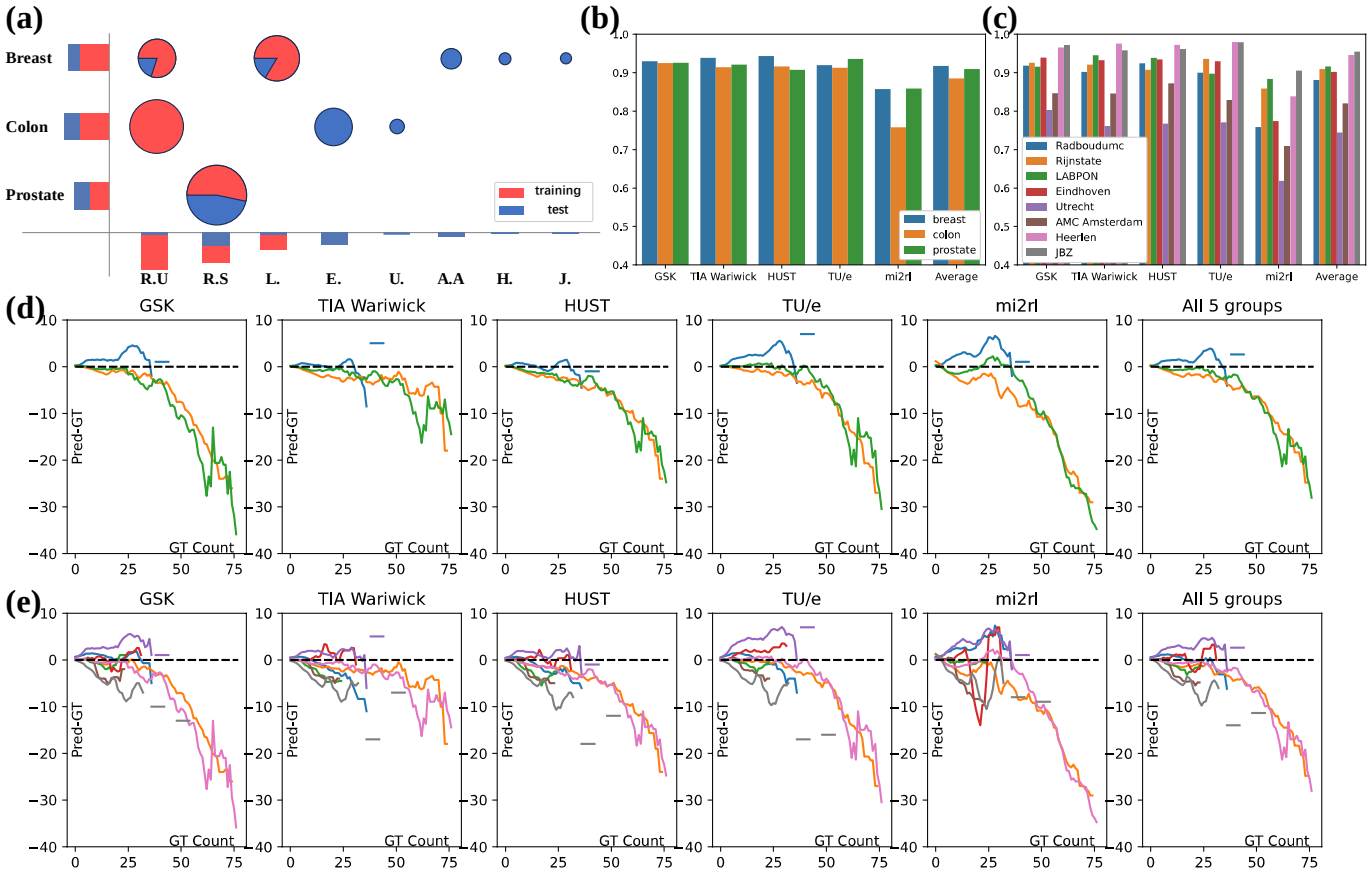

Fig. 6: Organ-level and institute-level data distribution, performance, and error trend. (a) sample number distribution; (b) QWK values on the organ level; (c) QWK values on the institute level; (d), (e) are the error trends, showing prediction error (pred-GT) regarding GT count in each subset, (d) and (e) are partitioned by organ and institute, respectively; the values are NaN-filtered and weighted by sample number with smoothing. R.U: Radboudumc; R.S: Rijnstate; L.: LABPON; E. Eindhoven; U.: Utrecht; A.A: AMC Amsterdam; H.: Heerlen; GT: ground-truth.

automatic methods. According to the recall matrices presented in Figure 5, five groups are also sharing consistency in terms of bin predictions. However, as shown in the first row, except for the TIA Warwick group, other groups prone to mis-predict label 6 as label 5 with different preferences for the two class.

The most challenging samples are related to background staining, artifacts, and brown debris, which often results in overestimations of positive cells. Additionally, the presence of weakly or partially stained cells also poses difficulties. Such difficult samples take at least 4% of the test set, where no teams gave correct prediction. Notably, this cannot be well addressed even using fully supervised methods [27]. So far, the best

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

This article has been accepted for publication in IEEE Journal of Biomedical and Health Informatics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2023.3327489

solution seems to identify regions of background staining and nonspecific staining as a pre-processing step under macro-level view [40]. Hard sample mining technique may also help [8].

Given the multi-organ and multi-cohort nature of LYSTO, the decline in model performance may also be due to domain shift caused by differences in data distribution[41]. To support this hypothesis, we visualized the distribution of data across organs, cohorts, and institutes, and evaluated the performance of five methods on the corresponding subsets of data in Figure 6 (a)~(c). One notable trend is the consistent decline in accuracy for all methods in the Utrecht (colon) cohort. Interestingly, the Eindhoven (colon) cohort, which is also an independent test set at the institute level, shows comparable QWK values for other subsets in the first four groups, whereas the mi2rl method exhibits a significant decrease in accuracy. These findings indicates that the performance in cross-domain scenario is influenced by both dataset characteristics and specific methods. In the breast subset, the Heerlen and JBZ cohorts can achieved even better results than the internal Radboudumc and LABPON sets in most cases, while the performance in the AMC Amsterdam cohort generally decreased. This once again confirms the complex effects caused by domain shift on the institute level.

In order to investigate whether there are common biases in different models, we further visualized the prediction error trends of each group at the subset level of organ and cohort in Figure 6 (d)~(e). It can be observed that all groups exhibit a noticeable negative bias, i.e., an underestimation of target count, especially when the ground-truth is large. This phenomenon is expected due to the relatively small number of labels with "$\geqslant$ 51". Increasing the number of such samples or generating training samples using carefully-designed generative methods may be potential solutions. We also discovered unique error trends in each data subset. For example, GSK, TU/e, and mi2rl exhibit a small but distinct positive bias in the breast cancer subset on the beginning part. Another interesting finding is the strikingly similar trends among all groups in the Utrecht cohort, despite that these methods were independently developed. This suggests that the cohort may induce a fixed-direction prediction bias that shares among various models.

The goal of LYSTO is to establish a lymphocyte evaluation pipeline with preferred generalization, where both internal and external performance are important. Therefore, our data arrangement may differ from other competitions. For example, data stained by Radboudumc, Rijnstate, and LABPON are present in both the training and test sets. A complete independent test set allows ones to find the best model for specific data. However, due to domain shift, the model may behave differently on a different test set. This phenomenon has been confirmed by comparing difference among institutes in LYSTO. Therefore, the most reasonable evaluation scheme remains an open question. In future research, it may be worth considering submitting an encapsulated training and evaluation framework to enable systematic evaluation on various subsets.

Performance metrics in LYSTO are computed based on local patches, which is different from whole-slide image used in practice. The external lung dataset and LYON dataset provide insights under broad field of view, as well as the generalization to a different organ and slides prepared by different centers. According to Table II, most of the methods got a QWK larger than 0.85. Mi2rl performance dropped, perhaps due to discrete predictions. Interestingly, the HUST method, which performed well on the LYSTO test set, shows a significant decrease in the lung dataset. We examined the predicted results and found frequent false positive on normal lung parenchyma and enlarged alveoli (data not shown). This indicates that even relatively simple IHC counting pipelines require thorough evaluation when used across tissue types.

The performance on LYON is defined by sensitivity for the convenient comparison with previous reader studies that involving a panel of pathologists [27]. As shown in Table III, most automated methods share similar performance patterns across bins. Compared to pathologists, these methods are prone to produce error in class ‘0’; however, when the cell count is extremely large (>=200), automatic methods give better prediction than human. According to the averaged sensitivities, four out of five methods achieved better performance than the pathologist panel. The performance of TU/e and HUST is even comparable with fully-supervised method[6], [27].

Based on LYSTO and post-event submission, we observed preference of using relatively simple models (such as VGG). These simple models have lower training and inference cost, with similar or even superior than complex counterparts. This perspective is supported by a series of studies [22], [23]. Despite recent advances in vision transformer (ViTs) and prompt learning, these methods have rarely been reported superior in IHC scoring tasks. Considering computational cost, simple models may be more suitable for IHC evaluations.

Numerous challenges have been hosted in the field of medical image analysis, which typically last for several months, allowing participants to iteratively update their methods [27], [42], [43]. Stemming from the cell counting problem, LYSTO attempted a novel challenge format, requiring participants to develop models and submit results within an extremely short timeframe. This brainstorming-style event encourages participants to focus on the problems and explore a wide range of possible solutions. For this, the data format and interface should be simplified and standardized as much as possible. The success of LYSTO demonstrates that with deep learning-based framework and well-prepared data, researchers can establish diagnostic models at a human expert level within a few hours.

Another major contribution of LYSTO is to promote lymphocyte assessment and computer vision research. LYSTO has already supported a series studies. These works mainly focus on lymphocyte IHC scoring and use detection models such as Faster R-CNN and Mask R-CNN [44]–[46]. Inspired by LYSTO, [47] explored an interactive annotation framework. Meanwhile, LYSTO can serve as benchmark dataset in computer vision field, for example, verifying novel group-invariance methods [48], [49].

As an early event, LYSTO has a clear limitation in that we did not require participants to provide source code or reusable models. While this is partly due to limited event time, it hinders us from conducting a more in-depth and detailed analysis of the results, as well as reusing these methods. Packaging the algorithms as Docker or other containers may serve as an effective solution to enhance algorithm availability and reproducibility[50].

We made the dataset, as well as the evaluation platform

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

This article has been accepted for publication in IEEE Journal of Biomedical and Health Informatics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2023.3327489

available on grand-challenge website. The dataset is also available via Zenodo (https://zenodo.org/record/3513571). In this way, we envision LYSTO as a potential future benchmark for development in computational pathology, easy to access and process.

VII. Conclusions

In this paper, we presented the summary of the Lymphocyte Assessment (LYSTO) Hackathon, which was held in conjunction with the 2019 Medical Image Computing and Computer Assisted Interventions (MICCAI) Conference. The aim of the hackathon was to develop automatic methods for immunohistochemistry quantification. We proposed the LYSTO dataset, which is composed of multi-center and multi-organ pathological images, as a reference to benchmark future computational pathology methods. Moreover, we left the LYSTO dataset as a long-lasting educational benchmark on https://lysto.grand-challenge.org/.

Acknowledgment

The authors would like to thank Nikki Wissink for her help in annotations of lymphocytes in lung tissue. Y. J. was funded by the China Scholarship Council for his internship in Radboud University Medical Center, and is funded by National Natural Science Funding of China (No. 62302228, 82330060). Jeroen van der Laak is a member of the advisory boards of Philips, The Netherlands and ContextVision, Sweden, and received research funding from Philips, The Netherlands, ContextVision, Sweden, and Sectra, Sweden in the last five years. Jeroen van der Laak is chief scientific officer of Aiosyn BV, Netherlands. Francesco Ciompi is shareholder of Aiosyn BV, Netherlands, and received consultancy fees from TRIBVN Healthcare, France. Z. L. received fundings from the National Natural Science Funding of China (No.61801491) and Natural Science Funding of Hunan Province (No.2019JJ50728).

LYSTO has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement no. 825292 (ExaMode project, http://www.examode.eu), and from the Alpe dHuZes / Dutch Cancer Society Fund, grant number KUN 2014-7032.

References

[1] H. Angell and J. Galon, "From the immune contexture to the Immunoscore: the role of prognostic and predictive immune markers in cancer," *Curr. Opin. Immunol.*, vol. 25, no. 2, pp. 261–267, Apr. 2013, doi: 10.1016/j.coi.2013.03.004.

[2] R. Salgado *et al.*, "The evaluation of tumor-infiltrating lymphocytes (TILs) in breast cancer: recommendations by an International TILs Working Group 2014," *Ann. Oncol.*, vol. 26, no. 2, pp. 259–271, 2015.

[3] M. Ono *et al.*, "Tumor-infiltrating lymphocytes are correlated with response to neoadjuvant chemotherapy in triple-negative breast cancer," *Breast Cancer Res. Treat.*, vol. 132, no. 3, pp. 793–805, 2012.

[4] J. Saltz *et al.*, "Spatial organization and molecular correlation of tumor-infiltrating lymphocytes using deep learning on pathology images," *Cell Rep.*, vol. 23, no. 1, pp. 181–193, 2018.

[5] W. H. Fridman, F. Pages, C. Sautes-Fridman, and J. Galon, "The immune contexture in human tumours: impact on clinical outcome," *Nat. Rev. Cancer*, vol. 12, no. 4, pp. 298–306, 2012.

[6] Z. Swiderska-Chadaj *et al.*, "Convolutional neural networks for lymphocyte detection in immunohistochemically stained whole-slide images," in *Proceedings of the 1st Conference on Medical Imaging with Deep Learning (MIDL 2018)*, 2018, pp. 1–12.

[7] M. Valkonen *et al.*, "Cytokeratin-Supervised Deep Learning for Automatic Recognition of Epithelial Cells in Breast Cancers Stained for ER, PR, and Ki-67," *IEEE Trans. Med. Imaging*, vol. 39, no. 2, pp. 534–542, Feb. 2020, doi: 10.1109/TMI.2019.2933656.

[8] D. Krijgsman *et al.*, "Quantitative Whole Slide Assessment of Tumor-Infiltrating CD8-Positive Lymphocytes in ER-Positive Breast Cancer in Relation to Clinical Outcome," *IEEE J. Biomed. Health Inform.*, vol. 25, no. 2, pp. 381–392, Feb. 2021, doi: 10.1109/JBHI.2020.3003475.

[9] R. Turkki, N. Linder, P. Kovanen, T. Pellinen, and J. Lundin, "Antibody-supervised deep learning for quantification of tumor-infiltrating immune cells in hematoxylin and eosin stained breast cancer samples," *J. Pathol. Inform.*, vol. 7, no. 1, p. 38, 2016, doi: 10.4103/2153-3539.189703.

[10] G. Corredor *et al.*, "Spatial Architecture and Arrangement of Tumor-Infiltrating Lymphocytes for Predicting Likelihood of Recurrence in Early-Stage Non–Small Cell Lung Cancer," *Clin. Cancer Res.*, vol. 25, no. 5, pp. 1526–1534, Mar. 2019, doi: 10.1158/1078-0432.CCR-18-2013.

[11] A. C. Ruifrok, D. A. Johnston, and others, "Quantification of histochemical staining by color deconvolution," *Anal. Quant. Cytol. Histol.*, vol. 23, no. 4, pp. 291–299, 2001.

[12] S. Chatterjee *et al.*, "Quantitative Immunohistochemical Analysis Reveals Association between Sodium Iodide Symporter and Estrogen Receptor Expression in Breast Cancer," *PLOS ONE*, vol. 8, no. 1, p. e54055, Jan. 2013, doi: 10.1371/journal.pone.0054055.

[13] R. S. Geread *et al.*, "IHC Color Histograms for Unsupervised Ki67 Proliferation Index Calculation," *Front. Bioeng. Biotechnol.*, vol. 7, p. 226, Oct. 2019, doi: 10.3389/fbioe.2019.00226.

[14] P. Bankhead *et al.*, "QuPath: Open source software for digital pathology image analysis," *Sci. Rep.*, vol. 7, no. 1, p. 16878, Dec. 2017, doi: 10.1038/s41598-017-17204-5.

[15] C.-C. Ko, C.-H. Lin, C.-H. Chuang, C.-Y. Chang, S.-H. Chang, and J.-H. Jiang, "A Whole Slide Ki-67 Proliferation Analysis System for Breast Carcinoma," in *2019 Twelfth International Conference on Ubi-Media Computing (Ubi-Media)*, Bali, Indonesia: IEEE, Aug. 2019, pp. 210–213. doi: 10.1109/Ubi-Media.2019.00048.

[16] N. J. Morriss, G. M. Conley, S. M. Ospina, W. P. Meehan III, J. Qiu, and R. Mannix, "Automated Quantification of Immunohistochemical Staining of Large Animal Brain Tissue Using QuPath Software," *Neuroscience*, vol. 429, pp. 235–244, Mar. 2020, doi: 10.1016/j.neuroscience.2020.01.006.

[17] S. Wu *et al.*, "The Role of Artificial Intelligence in Accurate Interpretation of HER2 Immunohistochemical Scores 0 and 1+ in Breast Cancer," *Mod. Pathol.*, vol. 36, no. 3, p. 100054, Mar. 2023, doi: 10.1016/j.modpat.2022.100054.

[18] M. Yue *et al.*, "Can AI-assisted microscope facilitate breast HER2 interpretation? A multi-institutional ring study," *Virchows Arch.*, vol. 479, no. 3, pp. 443–449, Sep. 2021, doi: 10.1007/s00428-021-03154-x.

[19] L. Cai *et al.*, "Improving Ki67 assessment concordance by the use of an artificial intelligence-empowered microscope: a multi-institutional ring study," *Histopathology*, vol. 79, no. 4, pp. 544–555, Oct. 2021, doi: 10.1111/his.14383.

[20] V. J. Tuominen, S. Ruotoistenmäki, A. Viitanen, M. Jumppanen, and J. Isola, "ImmunoRatio: a publicly available web application for quantitative image analysis of estrogen receptor (ER), progesterone receptor (PR), and Ki-67," *Breast Cancer Res.*, vol. 12, no. 4, p. R56, Aug. 2010, doi: 10.1186/bcr2615.

[21] M. Saha, C. Chakraborty, I. Arun, R. Ahmed, and S. Chatterjee, "An Advanced Deep Learning Approach for Ki-67 Stained Hotspot Detection and Proliferation Rate Scoring for Prognostic Evaluation of Breast Cancer," *Sci. Rep.*, vol. 7, no. 1, p. 3213, Dec. 2017, doi: 10.1038/s41598-017-03405-5.

[22] S. Tewary and S. Mukhopadhyay, "HER2 Molecular Marker Scoring Using Transfer Learning and Decision Level Fusion," *J. Digit. Imaging*, vol. 34, no. 3, pp. 667–677, Jun. 2021, doi: 10.1007/s10278-021-00442-5.

[23] S. Tewary and S. Mukhopadhyay, "AutoIHCNet: CNN architecture and decision fusion for automated HER2 scoring," *Appl. Soft Comput.*, vol. 119, p. 108572, Apr. 2022, doi: 10.1016/j.asoc.2022.108572.

[24] F. Negahbani *et al.*, "PathoNet introduced as a deep neural network backend for evaluation of Ki-67 and tumor-infiltrating lymphocytes in

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

This article has been accepted for publication in IEEE Journal of Biomedical and Health Informatics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2023.3327489

breast cancer," *Sci. Rep.*, vol. 11, no. 1, p. 8489, Dec. 2021, doi: 10.1038/s41598-021-86912-w.

[25] A. Kapil *et al.*, "Deep semi supervised generative learning for automated tumor proportion scoring on NSCLC tissue needle biopsies," *Sci. Rep.*, vol. 8, no. 1, pp. 1–10, 2018.

[26] M. van Rijthoven, Z. Swiderska-Chadaj, K. Seeliger, J. van der Laak, and F. Ciompi, "You only look on lymphocytes once," preprint, Apr. 2018. Accessed: Sep. 25, 2021. [Online]. Available: https://openreview.net/forum?id=S10IfW2oz

[27] Z. Swiderska-Chadaj *et al.*, "Learning to detect lymphocytes in immunohistochemistry with deep learning," *Med. Image Anal.*, vol. 58, p. 101547, 2019.

[28] F. Wilm *et al.*, "Pan-tumor T-lymphocyte detection using deep neural networks: Recommendations for transfer learning in immunohistochemistry," *J. Pathol. Inform.*, vol. 14, p. 100301, 2023, doi: 10.1016/j.jpi.2023.100301.

[29] M. M. K. Sarker *et al.*, "A Means of Assessing Deep Learning-Based Detection of ICOS Protein Expression in Colon Cancer," *Cancers*, vol. 13, no. 15, p. 3825, Jul. 2021, doi: 10.3390/cancers13153825.

[30] E. Garcia, R. Hermoza, C. B. Castanon, L. Cano, M. Castillo, and C. Castanneda, "Automatic Lymphocyte Detection on Gastric Cancer IHC Images Using Deep Learning," in *2017 IEEE 30th International Symposium on Computer-Based Medical Systems (CBMS)*, Thessaloniki: IEEE, Jun. 2017, pp. 200–204. doi: 10.1109/CBMS.2017.94.

[31] X. Wang *et al.*, "How can artificial intelligence models assist PD-L1 expression scoring in breast cancer: results of multi-institutional ring studies," *Npj Breast Cancer*, vol. 7, no. 1, p. 61, Dec. 2021, doi: 10.1038/s41523-021-00268-y.

[32] T. Qaiser *et al.*, "HER2 challenge contest: a detailed assessment of automated HER2 scoring algorithms in whole slide images of breast cancer tissues," *Histopathology*, vol. 72, no. 2, pp. 227–238, Jan. 2018, doi: 10.1111/his.13333.

[33] M. C. A. Balkenhol *et al.*, "Deep learning and manual assessment show that the absolute mitotic count does not contain prognostic information in triple negative breast cancer," *Cell. Oncol.*, vol. 42, no. 4, pp. 555–569, Aug. 2019, doi: 10.1007/s13402-019-00445-z.

[34] G. Litjens *et al.*, "1399 H&E-stained sentinel lymph node sections of breast cancer patients: the CAMELYON dataset," *GigaScience*, vol. 7, no. 6, p. giy065, 2018.

[35] L. N. Smith and N. Topin, "Super-convergence: Very fast training of neural networks using large learning rates," in *Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications*, International Society for Optics and Photonics, 2019, p. 1100612.

[36] S. Graham *et al.*, "Hover-net: simultaneous segmentation and classification of nuclei in multi-tissue histology images," *Med. Image Anal.*, vol. 58, p. 101563, Dec. 2019, doi: 10.1016/j.media.2019.101563.

[37] O. Sagi and L. Rokach, "Ensemble learning: A survey," *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.*, vol. 8, no. 4, p. e1249, 2018.

[38] R. Mormont, P. Geurts, and R. Maree, "Multi-Task Pre-Training of Deep Neural Networks for Digital Pathology," *IEEE J. Biomed. Health Inform.*, vol. 25, no. 2, pp. 412–421, Feb. 2021, doi: 10.1109/JBHI.2020.2992878.

[39] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?," *ArXiv Prepr. ArXiv14111792*, 2014.

[40] A. Abreu *et al.*, "Ensemble Of Neural Networks For High Endothelial Venules Detection In Meca-79 Immunohistochemistry Images," in *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, Venice, Italy: IEEE, Apr. 2019, pp. 938–942. doi: 10.1109/ISBI.2019.8759578.

[41] K. Stacke, G. Eilertsen, J. Unger, and C. Lundstrom, "Measuring domain shift for deep learning in histopathology," *IEEE J. Biomed. Health Inform.*, vol. 25, no. 2, pp. 325–336, 2020, doi: 10.1109/JBHI.2020.3032060.

[42] K. Sirinukunwattana *et al.*, "Gland segmentation in colon histology images: The glas challenge contest," *Med. Image Anal.*, vol. 35, pp. 489–502, Jan. 2017, doi: 10.1016/j.media.2016.08.008.

[43] B. E. Bejnordi *et al.*, "Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer," *JAMA*, vol. 318, no. 22, pp. 2199–2210, 2017, doi: 10.1001/jama.2017.14585.

[44] I. Keren Evangeline, J. Glory Precious, N. Pazhanivel, and S. P. Angeline Kirubha, "Automatic Detection and Counting of Lymphocytes from Immunohistochemistry Cancer Images Using Deep Learning," *J. Med. Biol. Eng.*, vol. 40, no. 5, pp. 735–747, Oct. 2020, doi: 10.1007/s40846-020-00545-4.

[45] M. M. Zafar *et al.*, "Detection of tumour infiltrating lymphocytes in CD3 and CD8 stained histopathological images using a two-phase deep CNN," *Photodiagnosis Photodyn. Ther.*, vol. 37, p. 102676, Mar. 2022, doi: 10.1016/j.pdpdt.2021.102676.

[46] Z. Rauf, A. Sohail, S. H. Khan, A. Khan, J. Gwak, and M. Maqbool, "Attention-guided multi-scale deep object detection framework for lymphocyte analysis in IHC histological images," *Microscopy*, vol. 72, no. 1, pp. 27–42, 2023.

[47] M. M. Zafar, Z. Rauf, A. Sohail, and A. Khan, "Lymphocyte Annotator: CD3 $^{+}$ and CD8 $^{+}$ IHC Stained Patch Image Annotation Tool," in *2020 International Symposium on Recent Advances in Electrical Engineering & Computer Sciences (RAEE & CS)*, Islamabad, Pakistan: IEEE, Oct. 2020, pp. 1–6. doi: 10.1109/RAEECS50817.2020.9265757.

[48] J. Birrell, M. A. Katsoulakis, L. Rey-Bellet, and W. Zhu, "Structure-preserving GANs." arXiv, Jun. 17, 2022. Accessed: May 11, 2023. [Online]. Available: http://arxiv.org/abs/2202.01129

[49] N. Dey, A. Chen, and S. Ghafurian, "Group Equivariant Generative Adversarial Networks." arXiv, Mar. 30, 2021. Accessed: May 11, 2023. [Online]. Available: http://arxiv.org/abs/2005.01683

[50] Q. Da *et al.*, "DigestPath: A benchmark dataset with challenge review for the pathological detection and segmentation of digestive-system," *Med. Image Anal.*, vol. 80, p. 102485, Aug. 2022, doi: 10.1016/j.media.2022.102485.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/