# A Multi-organ Nucleus Segmentation Challenge

Neeraj Kumar*, Ruchika Verma, Deepak Anand, Yanning Zhou, Omer Fahri Onder, Efstratios Tsougenis, Hao Chen, Pheng-Ann Heng, Jiahui Li, Zhiqiang Hu, Yunzhi Wang, Navid Alemi Koohbanani, Mostafa Jahanifar, Neda Zamani Tajeddin, Ali Gooya, Nasir Rajpoot, Xuhua Ren, Sihang Zhou, Qian Wang, Dinggang Shen, Cheng-Kun Yang, Chi-Hung Weng, Wei-Hsiang Yu, Chao-Yuan Yeh, Shuang Yang, Shuoyu Xu, Pak Hei Yeung, Peng Sun, Amirreza Mahbod, Gerald Schaefer, Isabella Ellinger, Rupert Ecker, Orjan Smedby, Chunliang Wang, Benjamin Chidester, That-Vinh Ton, Minh-Triet Tran, Jian Ma, Minh N. Do, Simon Graham, Quoc Dang Vu, Jin Tae Kwak, Akshaykumar Gunda, Raviteja Chunduri, Corey Hu, Xiaoyang Zhou, Dariush Lotfi, Reza Safdari, Antanas Kascenas, Alison O'Neil, Dennis Eschweiler, Johannes Stegmaier, Yanping Cui, Baocai Yin, Kailin Chen, Xinmei Tian, Philipp Gruening, Erhardt Barth, Elad Arbel, Itay Remer, Amir Ben-Dor, Ekaterina Sirazitdinova, Matthias Kohl, Stefan Braunewell, Yuexiang Li, Xinpeng Xie, Linlin Shen, Jun Ma, Krishanu Das Baksi, Mohammad Azam Khan, Jaegul Choo, Adrián Colomer, Valery Naranjo, Linmin Pei, Khan M. Iftekharuddin, Kaushiki Roy, Debotosh Bhattacharjee, Anibal Pedraza, Maria Gloria Bueno, Sabarinathan Devanathan, Saravanan Radhakrishnan, Praveen Koduganty, Zihan Wu, Guanyu Cai, Xiaojie Liu, Yuqin Wang, and Amit Sethi

***Abstract*—Generalized nucleus segmentation techniques can contribute greatly to reducing the time to develop and validate visual biomarkers for new digital pathology datasets. We summarize the results of MoNuSeg 2018 Challenge whose objective was to develop generalizable nuclei segmentation techniques in digital pathology. The challenge was an official satellite event of the MICCAI 2018 conference in which 32 teams with more than 80 participants from geographically diverse institutes participated. Contestants were given a training set with 30 images from seven organs with annotations of 21,623 individual nuclei. A test dataset with 14 images taken from seven organs, including two organs that did not appear in the training set was released without annotations. Entries were evaluated based on average aggregated Jaccard index (AJI) on the test set to prioritize accurate instance segmentation as opposed to mere semantic segmentation. More than half the teams that completed the challenge outperformed a previous baseline [1]. Among the trends observed that contributed to increased accuracy were the use of color normalization as well as heavy data augmentation. Additionally, fully convolutional networks inspired by variants of U-Net [2], FCN [3], and Mask-RCNN [4] were popularly used, typically based on ResNet [5] or VGG [6] base architectures. Watershed segmentation on predicted semantic segmentation maps was a popular post-processing strategy. Several of the top techniques compared favorably to an individual human annotator and can be used with confidence for nuclear morphometrics.**

***Index Terms*—Multi-organ, nucleus segmentation, digital pathology, instance segmentation, aggregated Jaccard index.**

## I. Introduction

Examination of H&E stained tissue under a microscope remains the mainstay of pathology. The popularity of H&E is due to its low cost and ability to reveal tissue structure and nuclear morphology, which is sufficient for primary diagnosis of several diseases including many cancers. Nuclear shapes and spatial arrangements often form the basis of the examination of H&E stained tissue sections. For example, grading of various types of cancer and risk stratification of patients is usually done by examining different types of nuclei on a tissue slide [7], [8]. Nuclear morphometric features and appearance including the color of their surrounding cytoplasm also helps in identifying various types of cells such as epithelial (glandular), stromal, or inflammatory, which in turn give an idea of the glandular structure and disease presentation at low power [7]–[10]. Segmentation of nuclei accurately in H&E images therefore has high utility in digital pathology.

However, nucleus segmentation algorithms that work well on one dataset can perform poorly on a different dataset. There is far too much variation in the appearance of nuclei and their surroundings by organs, disease conditions, and even digital scanner brands or histology technicians. Examples of such variations are shown in Figure 1, along with the problems of some common segmentation algorithms such as Otsu thresholding [11], marker controlled watershed segmentation [12]–[14] or open-source packages like Fiji [15] and Cell Profiler [16]. Segmentation based on machine learning should be able to do a better job, but that makes designing and refining nucleus segmentation algorithms for a new study a tedious task because annotations of thousands of nuclei are needed to train such segmentation models on datasets of interest. Algorithms that generalize to new datasets and organs that were not seen during training can reduce this effort substantially and contribute to rapid experimentation with new phenotypical (visual) biomarkers.

Until recently, one of the major challenges in training generalized nucleus segmentation models has been the unavailability of large multi-organ datasets with annotated nuclei. In 2017 Kumar *et al*. [1] released a dataset with more than 21,000

N. Kumar, R. Verma, D. Anand and A. Sethi co-organized the challenge; all others contributed the results of their algorithms. Due to space constraints, funding information and author affiliations for this work appear in the acknowledgement section. Asterisk indicates the corresponding author. Address all correspondence to: neeraj.kumar.iitg@gmail.com, and asethi@iitb.ac.in

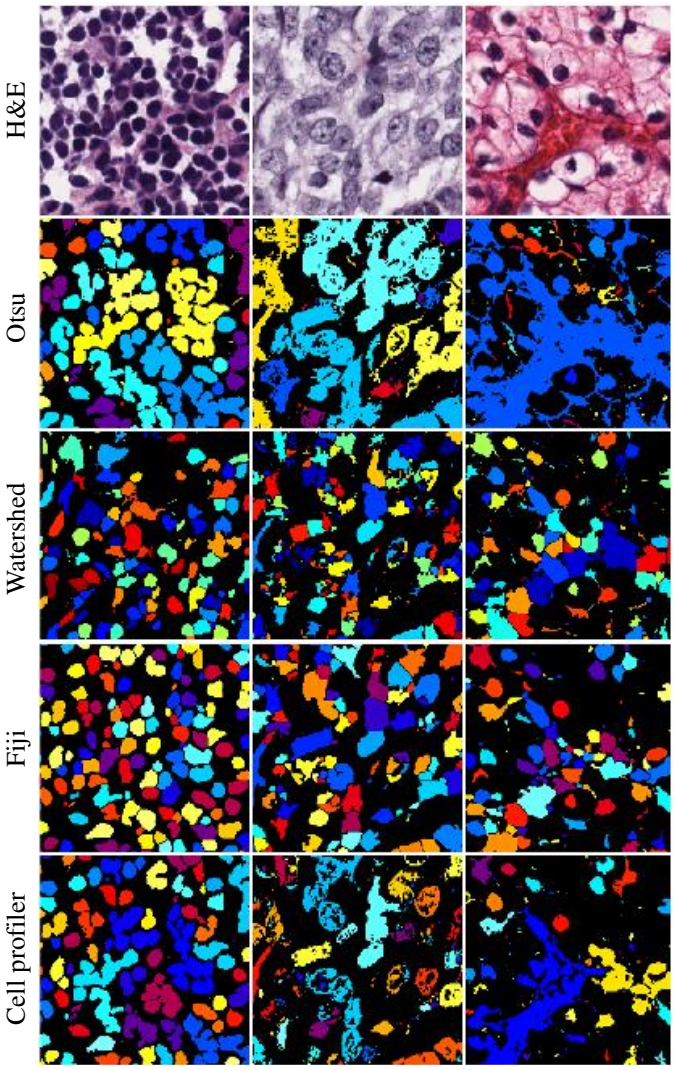


Fig. 1: Nucleus segmentation challenges: Original H&E images show crowded and chromatin-sparse nuclei with color variation across tissue slides. Otsu thresholding [11] and Cell Profiler [16] gives merged nuclei (under-segmentation). Marker controlled watershed segmentation [12] and Fiji [15] produces fragmented nuclei (over-segmentation). Segmented nuclei instances are shown in different colors in rows 2-5.

hand-annotated nuclei in H&E stained tissue images acquired at the commonly used 40x magnification, sourced from seven organs and multiple hospitals in The Cancer Genome Atlas (TCGA) [17]. Kumar *et al*. also introduced a metric called Aggregated Jaccard Index (AJI) that is more appropriate to evaluate algorithms for this instance segmentation problem as opposed to other popular metrics such as Dice coefficient, which are more suited for semantic segmentation problems. This is because nucleus segmentation algorithms should not only tell the difference between nuclear and non-nuclear pixels, but they should also be able to tell pixels belonging to two nuclei apart that touch or overlap with each other. Additionally, they had released a trained model that performed reasonably well on unseen organs from the test subset of images.

We organized the Multi-organ nucleus segmentation (MoNuSeg) Challenge at MICCAI 2018 to build upon Kumar *et al*.'s work by enlarging the dataset and by encouraging others to introduce new techniques for generalized nucleus segmentation. The participation was wide and several of participants outperformed the previous benchmark [1] by a significant margin. In this paper we describe in detail the objectives of the competition, the released dataset, and the emerging trends of techniques that performed well on the challenge task. We hope that the algorithms described on the challenge webpage [18] will be of use to the computational pathology research community.

The rest of the paper is organized as follows. We describe the prior work on nucleus segmentation and dataset creation in Section II. We describe the dataset and competition rules in Section III. We present an organized summary of the techniques used by the challenge participants in Section IV. Finally, we discuss emerging trends in nucleus segmentation techniques in Section V.

## II. Background and Prior Work

In this section we describe the importance of H&E stained images in histopathology and provide details of some previous notable techniques and datasets for nucleus segmentation from H&E stained images.

### A. Hematoxylin and eosin (H&E) stained images

Pathologists usually observe tissue slides under a microscope at a specific resolution (ranging between 5x and 40x) to report their diagnoses including tumor grade, extent of spread, surgical margin, etc. Their assessment is primarily based on the appearance, size, shape, color and crowding of various nuclei (and glands) in epithelium and stroma. Stains are used to enhance the contrast between these tissue components to help a pathologist looking for specific nuclei and gland features. The combination of hematoxylin and eosin (H&E) is a frequently-used, universal, and inexpensive staining scheme for general contrast enhancement of histologic structures of a tissue. Hematoxylin renders the nuclei dark blueish purple and the epithelium light purple, while eosin renders the stroma pink. Compared to the general use of H&E, immunohistochemical staining is more specialized as it targets proteins specific to certain disease states for visual identification.

With the advent of high resolution cameras mounted on microscopes, and more importantly, digital whole slide scanners, it is now possible to acquire whole slide images (WSIs) of the tissue sections for computer assisted diagnosis (CAD). However, the development of CAD systems requires automated extraction of rich information encoded in the pixels of WSIs. Recently, computer based assessment of tissue images has been used for tumor molecular sub-type detection [19], mortality or recurrence prediction [9], [20], and treatment effectiveness prediction [10]. Notably, nucleus detection and segmentation is often a first step for several such CAD systems that rely on nuclear morphometrics for disease state stratification and predictive modelling. Therefore, MoNuSeg 2018 focused on crowdsourcing *techniques* for nucleus segmentation in H&E stained images captured at 40x resolution.

### B. Nucleus segmentation techniques

Prior to the advent of deep learning, approaches to segment nuclei relied on watershed segmentation, morphological operations – such as erosion, dilation, opening and closing – color-based thresholding, and variants of active contours [12], [13], [21]–[23]. These techniques were often complemented with a collection of pre-processing methods, such as contrast enhancement and deblurring to improve the 'image quality'. Additionally, several post-processing techniques, such as hole filling, noise removal, graph-cuts, etc., were also used to refine the outputs of the segmentation algorithms. However, these approaches do not generalize well across a wide spectrum of tissue images due to reasons such as (a) variations in nuclei morphologies of various organs and tissue types, (b) inter- and intra-nuclei color variations in crowded and chromatin-sparse nuclei, and (c) diversity in the quality of tissue images owing to the differences in image acquisition equipment and slide preparation protocols across hospitals and clinics.

There have been tremendous advances in the recent years to develop learning-based nucleus segmentation methods to advance the state-of-the-art. Instead of relying on pre-determined algorithms for segmentation, machine learning methods derive data driven algorithms that are trained in a supervised manner based on annotations of nuclear and non-nuclear pixels. This allows them to concentrate on relative differences between nuclear and non-nuclear pixels and their surrounding patches and overcome the aforementioned sources of intra-class variations for better generalized segmentation. The use of learning based approaches started with the extraction of hand-crafted local features based on color and spatial filtering that were fed to traditional learning-based models such as random forests, support vector machines, etc. to segment nuclei and non-nuclei regions [24]–[26]. The selection of features is dependent on domain knowledge and trial-and-error for improving nucleus segmentation performance, and yet it is difficult to detect all nuclei with diverse appearances and crowding patterns.

To circumvent the constraints of hand-crafted features, representation learning algorithms, popularly known as deep learning techniques, have recently emerged. These methods – specifically the ones using convolutional neural networks (CNNs) – have outperformed previous techniques in nucleus detection and segmentation tasks by significant margins [1], [27]–[30]. To use deep learning, the problem is often cast as one of semantic segmentation wherein a two-class probability map for nuclear and non-nuclear regions is usually computed. After semantic segmentation, sophisticated post-processing methods – such as graph partitioning [27], or the computation of distance transform of the nuclear map followed by H-minima transform and region growing [28] – are often used to obtain final nuclei shapes with the desired separation of touching and overlapping nuclei. Semantic segmentation of third class of pixels – those on the nuclear boundaries including that between two touching nuclei – has also been proposed to exclusively refine the separation between the segmented touching and overlapping nuclei [1]. Deep generative models have also been used for accurate nuclei segmentation [31]. More recently, nucleus segmentation problem has been formulated as a regression task to predict a distance map with respect to centroids or boundaries of nuclei using fully convolutional networks (FCNs) to achieve both segmentation and computational performance gains over previous deep learning based approaches [30]. More comprehensive reviews of state-of-the-art nucleus segmentation algorithms can be found in [32] and [33].

One of the major barriers in out of the box (without re-training) application of state-of-the art deep learning based nucleus segmentation algorithms was the lack of publicly available source codes and trained models by previously published techniques until Kumar *et al*. [1] and Naylor *et al*. [30] released their source codes. The other major barrier was the lack of publicly available annotated datasets for benchmarking, which we address next.

### C. Nucleus segmentation datasets

The success of machine learning and the development of state-of-the art deep learning algorithms in computer vision can be attributed to the healthy competition enabled by publicly available consumer photography datasets such as ImageNet [34] and COCO [35] for object recognition in images. Unfortunately, we do not see similar progress in digital pathology image analysis as there is dearth of labeled and annotated datasets for solving various tasks of pathologist's interest. For example, CAMELYON dataset [36], which is one of the largest histopathology classification dataset, has 1,399 images, while ImageNet [34] has 14 million images. Similarly, CheXpert [37], which is one of the largest medical image segmentation datasets has only $224,316$ images. This is because labeling and annotating pathology images require expert knowledge and diligent work. However, there have been a few recent efforts dedicated to the release of hand-annotated H&E stained tissue slide images for nucleus segmentation as summarized in Table I. These datasets can also be downloaded from the challenge webpage [18]. Please note that we have not included datasets where the nuclei were annotated for detection alone in Table I because these cannot be used for the segmentation task. We also excluded datasets annotated for other specific objectives such as gland segmentation, mitosis detection, epithelial segmentation, and tumor type classification, as opposed to generalized nucleus segmentation. Most of the datasets listed in Table I focus on a specific organ with the exception of Kumar *et al*. [1] and Wienert *et al*. [23].

## III. Dataset and Competition Rules

The objective of MoNuSeg 2018 was to encourage the development of learning based generalized nucleus segmentation techniques that work right out of the box (without re-training) on a diverse set of H&E-stained tissue images. The images therefore spanned a range of patients, organs, disease states, and sourcing hospitals with potentially different slide preparation and image acquisition methods. Training and testing datasets were carefully curated and the competition rules were crafted in accordance with these objectives.

[1]Only annotations verified by a pathologist were considered.

TABLE I: Publicly available H&E stained tissue image datasets annotated for nucleus segmentation

| Dataset | Image Size | Images | Nuclei | # Organs | Annotation type |
|---|---|---|---|---|---|
| Kumar *et al*. [1] | $1000 \times 1000$ | 30 | 21,623 | 7 | Individual nuclei boundary coordinates |
| Janowczyk *et al*. [38] | $2000 \times 2000$ | 143 | 12,000 | 1 (breast) | Binary foreground mask |
| Wienert *et al*. [23] | $600 \times 600$ | 36 | 7,931 | 5 | Individual nuclei boundary coordinates |
| Naylor *et al*. [30] | $512 \times 512$ | 50 | 4,022 | 1 (breast) | Binary foreground mask |
| Irshad *et al*.[1] [39] | $400 \times 400$ | 63 | 2,532 | 1 (kidney) | Binary foreground mask |
| Gelasca *et al*. [40] | $896 \times 768$ $(768 \times 512)$ | 50 | 1,895 | 1 (breast) | Binary foreground mask |

TABLE II: MoNuSeg 2018 training and testing datset composition.

| Data subset ↓ | Nuclei | Images | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | Total | Total | Breast | Liver | Kidney | Prostate | Bladder | Colon | Stomach | Lung | Brain |
| Training set | 21,623 | 30 | 6 | 6 | 6 | 6 | 2 | 2 | 2 | – | – |
| Testing set | 7,223 | 14 | 2 | – | 3 | 2 | 2 | 1 | – | 2 | 2 |
| **Total** | **28,846** | **44** | **8** | **6** | **9** | **8** | **4** | **3** | **2** | **2** | **2** |

### *A. Training dataset*

The training data of MoNuSeg 2018 was the same as that released previously by Kumar *et al*. [1], which comprised 30 tissues images, each of size $1000 \times 1000$, containing $21,623$ hand-annotated nuclear boundaries. Each $1000 \times 1000$ image in this dataset was extracted from a separate whole slide image (WSI) (scanned at 40x) of an individual patient downloaded from TCGA [17]. The dataset represented 7 different organs *viz.*, breast, liver, kidney, prostate, bladder, colon and stomach, and included both benign and diseased tissue samples to ensure diversity of nuclear appearances. Furthermore, the training images came from 18 different hospitals, which introduced another source of appearance variation due to the differences in the staining practices and image acquisition equipments (scanners) across labs. Representative $1000 \times 1000$ sub-images from regions dense in nuclei were extracted from patient WSIs to reduce the computational burden of processing WSIs and increase participation. Only one crop per WSI and patient was included in the dataset to ensure diversity. The distribution of training images across organs is shown in Table II while patient and hospital details are available on the challenge webpage [18].

Both epithelial and stromal nuclei were manually annotated in the $1000 \times 1000$ sub-images using Aperio ImageScope®. Annotations were performed on a 25" monitor with a 200x digital magnification such that each image pixel occupied $5 \times 5$ screen pixels to ensure clear visibility for annotating nuclear boundaries with a laser mouse. For overlapping nuclei, each multi-nuclear pixel was assigned to the nucleus that appeared to be on top in the 3-D structure. The annotators were engineering students and the quality control was performed by an expert pathologist with years of experience in analyzing tissue sections. Specifically, each H&E image was included in a PowerPoint® (Microsoft, Redmond WA, USA) slide at 300 dots per inch, along with the annotated boundaries overlaid in bright green. The slides were examined by a pathologist on 25" monitor to point out missed nuclei, false nuclei, and nuclei with wrong boundaries. For each image, the numbers of each type of error was summed up and divided by the number of annotated nuclei to assess the quality of annotations. As shown in Supplementary Table S2, the error rate for each organ was smaller than 1%. The images and XML files containing pixel coordinates of the annotated nuclear boundaries were released for public use by [1]. The reasons that make this dataset ideal for training a generalized nucleus segmentation model are as follows:

1) It is the largest repository of hand annotated nuclei which aptly represents a miscellany of nuclei shapes, and sizes across multiple organs, disease states and patients. The inclusion of tissue sections from 18 hospitals further augments the richness of this dataset. From Table I, the only multi-organ alternative to it is Weinert *et al*. [23]. However, Weinert *et al*. [23] contains tissues from lesser number of organs captured in a single hospital with a single scanner.
2) It extracted only one sub-image of $1000 \times 1000$ pixels per patient to maximize nuclear appearance variation. Other datasets mentioned in Table I extracted multiple sub-images from each patient and are thus limited in representing nuclear appearance diversity. For example, WSIs of only 10 and 11 patients were used in Irshad *et al*. [39] and Naylor *et al*. [30], respectively.
3) It provided coordinates of annotated nuclear boundaries in an XML format instead of binary masks. This is crucial for learning to separate touching and overlapping nuclei in any automatic nucleus segmentation algorithm. This helped several participants of MoNuSeg 2018 whose nucleus segmentation algorithms explicitly learned to recognize nuclear boundaries in addition to the usual foreground (nuclei pixels) and background classes (non-nuclei pixels).
4) It publicly released the source code of their generalized nucleus segmentation algorithm to catalyze natural competition among a newer generation of automatic nucleus segmentation algorithms.

### *B. Testing dataset*

A new testing set comprising 14 images, each of size $1000 \times 1000$ pixels, spanning 7 organs (*viz.* kidney, lung, colon, breast, bladder, prostate, brain), several disease states (benign and tumors at different stages), and approximately 7,223 annotated nuclei was prepared in the same manner as used for preparing the training data. As shown in Table II, lung and brain tissue images were exclusive to the test set which made it more challenging. More details about the test set are available in the "supplementary material" tab of the

challenge webpage [18]. The annotations of the test set were not released to the participants. To formally conclude the challenge, with this paper, we are releasing the test annotations on the challenge webpage [18] to facilitate future research in the development of generalized nucleus segmentation algorithms.

### C. Competition metric and Results

Competitors were evaluated only once on the test set. Their latest submission before the deadline was considered as the final submission for evaluation. Average aggregated Jaccard Index (AJI) was used as the metric to evaluate nucleus segmentation performance of the competing algorithms because of its established advantages over other segmentation metrics [1], [30]. The value of AJI ranges between 0 to 1 (higher is better). Computing AJI involves matching every ground truth nuclei to one detected nuclei by maximizing the Jaccard index. The AJI is then equal to the ratio of the sums of the cardinals of intersection and union of these matched ground truth and predicted nuclei. Additionally, all detected components that are not matched are added to the denominator. We reproduce Algorithm 1 detailing AJI computation from [1] with permission. The code for computing AJI is available on the challenge webpage [18].

**Algorithm 1** Aggregated Jaccard index (AJI)

**Input:** A set of images with a combined set of annotated nuclei $G_i$ indexed by $i$, and a segmented set of nuclei $S_k$ indexed by $k$.
**Output:** Aggregated Jaccard Index $A$.

Initialize overall correct and union pixel counts: $C \leftarrow 0; U \leftarrow 0$
**for** Each ground truth nucleus $G_i$ **do**
  $j \leftarrow \arg\max_k(|G_i \cap S_k|/|G_i \cup S_k|)$
  Update pixel counts: $C \leftarrow C + |G_i \cap S_j|; U \leftarrow U + |G_i \cup S_j|$
  Mark $S_j$ *used*
**end for**
**for** Each segmented nucleus $S_j$ **do**
  If $S_k$ is not *used* then $U \leftarrow U + |S_k|$
**end for**
$A \leftarrow C/U$

Participants were asked to submit 14 segmentation output files (one for each of the 14 test images) to the challenge organizers. For each participant's submission, the organizers then computed 14 AJIs (one for each test image) as per Algorithm 1. If a participant did not submit the results for a particular testing image then AJI value of zero was assigned for that particular image to that participant. The organizers then computed the average AJI (a-AJI) for each participant by averaging image level AJIs across 14 test images. The participants were then ranked in the descending order of a-AJI to obtain the final leaderboard shown in Table III.

Table III also includes the 95% confidence intervals (CIs) around each participant's a-AJI. It is evident that the confidence intervals of the top five techniques exclude a-AJI of the lower ranked techniques. To further assess the overall a-AJI based ranking scheme, we also computed organ level a-AJI (and confidence intervals), for each participant, by averaging image level AJIs across the number of images that belonged to a specific organ, as shown in Supplementary Table S3. From Supplementary Table S3, it is evident that (a) the top five techniques perform better than other techniques for each organ as well, (b) the organ is a larger contributor to the variability in performance among the top five techniques than the technique itself, and (c) techniques with a higher overall a-AJI perform better for more organs even among the top five techniques. Specifically, for instance, (a) no technique that is not among the top-five overall breaks into the top-five for more than two organs, (b) breast cancer images had AJIs that were lower by about 0.063 to 0.085 compared to those for bladder for the top-five techniques, and (c) the overall top-ranked technique is also the top-ranked one for all but one organ.

## IV. Summary of Segmentation Techniques

In this section we present a summary of the techniques used by 32 teams who successfully completed the challenge. We describe the trends observed in pre-processing, data augmentation, modeling, task specification, optimization, and post-processing techniques used by the teams. Specific details of all algorithms are provided in the respective manuscripts submitted by participants as per challenge policies and are available at challenge webpage [18] under "manuscripts" tab.

### A. Pre-processing and data augmentation

Pre-processing techniques reduce unwanted variations among input images – from both the training and testing sets – so that the test data distribution is not very different from the training data distribution, by projecting both to the same low-dimensional manifold. On the other hand, data augmentation techniques increase the training data set size by introducing controlled random variations with the hope of creating a training data distribution that covers most of the test data distribution. There are several ways in which the participants altered the given images and their ground truth masks before passing them to the segmentation learning systems in order to increase test accuracy. We summarize some of the interesting trends observed in this challenge. These results are also summarized in Table III.

*1) Color and intensity normalization:* Among the data pre-processing techniques, color and intensity transformations were the most common. Approximately half the teams used color normalization techniques that were specifically developed for pathology images to reduce unwanted color variations between training and testing data. Structure Preserving Color Normalization (SPCN) by Vahadane *et al*. [41] was used by ten teams due to its demonstrated performance and code availability. Another seven teams used Mecenko *et al*.'s color normalization scheme [42], out of which one used this technique in combination with another technique by Reinhard *et al*. [43].

Pixel intensity and RGB color transformations that are unspecific to pathology were also used by approximately half of the teams. Most popular among this class of techniques were channel-wise mean subtraction, variance normalization (unit variance), and pixel-value range standardization. Six teams also used either contrast enhancement (or histogram equalization), among which CLAHE [44] was the most commonly used technique.

Among the unique techniques, one team used image sharpening to remove unwanted variations between training and testing data, one team concatenated HSV and L channels (of L, a*, b* color space) to the RGB channels, and one team used only the blue channel after color normalization of the RGB images.

*2) Data augmentation:* Among data augmentation techniques, geometric transformations of the image grid were the most common. For example, rigid transformations of the images – such as rotation (especially, by multiples of 90 degrees) and/or flipping – were used by all but four teams to increase the size of the training data. However, as can be seen in Table III, all of the top twelve teams by a-AJI also augmented the training set using affine transformations, while only five teams below that used this type of augmentation. Another transformation used by the participants was elastic deformation, but it was not very popular among the contestants due to the marginal gain it might afford over an affine transform, while being more complicated to implement. Another geometric transformation is image scaling, which was used by nine contestants.

Another popular set of augmentation techniques involve changing the pixel values while leaving the geometric structure intact. The most popular among these techniques was the addition of white Gaussian noise, which was used by several of the top performing teams. Another popular technique is color jitter or random HSV shifts, which was used by nine of the top twelve teams. Color jitter is opposite in spirit to color normalization in that it is used to present more color variations of the same input geometric structure to the learning machine with the hope that it will learn to focus on the geometric structure as opposed to the color of nuclei, which may vary between training and testing data sets. Random intensity (brightness) shifts were used by fewer participants, as were blurring by isotropic Gaussian filters of random widths and random image sharpening.

One interesting data augmentation technique that was used by team *CMU-UIUC* involved extracting the nuclei, augment them in-place, filling the holes in the background, and then pasting the nuclei back on to the reconstructed background.

### B. Specification of the learning task

The challenge of nucleus segmentation can be split into two tasks: distinguishing between nuclear and non-nuclear pixels (semantic segmentation) and separating touching nuclei (instance segmentation). The following were three principal types of outputs that the contestants produced using deep learning to meet these two challenges:

1) **Binary class probability** maps distinguish between pixels that belong to the core of any nucleus versus those that do not. The process of not including the outer periphery of the nuclei into the foreground class helps separate touching nuclei. The lost nuclear territory can later be gained back during post-processing.
2) **Ternary class probability** map distinguishes between nuclear core, non-nuclear, and nuclear boundary pixels. Nuclear pixels that are on a shared boundary of two touching nuclei are considered to belong to the third class, which has been shown to be useful in separating touching nuclei [1].
3) **Distance map** estimates how far a nuclear pixel is from the centroid of a nucleus. Such a map can also distinguish between nuclear and non-nuclear pixels by assigning a fixed value to the latter, such as 0. This is a per-pixel regression problem while the previous two are classification problems. A variant of this distance map is to predict the distance from the boundary of the nucleus.

Most teams trained their models to predict variants of one or more of the three types of maps described above. One interesting departure from these three tasks was by Canon Medical Research Europe who predicted a five-class probability map – one for nuclear pixels, and the other four for their probability of belonging to one of the four Cartesian quadrants of a nucleus in order to separate touching nuclei.

### C. Model architectures

All participants used deep convolutional neural networks. Twenty one teams used variants of U-Net [2], of which the original U-Net architecture was used by 11 teams while six teams used base architectures inspired by VGGNet [6], and another 11 teams used architectures inspired by either MRCNN [4], FCN [3], DenseNet [45], or ResNet [5] with different depths. Eight teams used Mask Region with CNN features (MRCNN) [4] as the primary models (of which, two also used U-Net), and two used FCN [3] (of which one also used U-Net). Among the remaining, four teams used their own custom models and architectures, and one each used VGGNet [6], Deep Layer Aggregation [46], PANet [47], and TernausNet [48]. A few teams used multiple architectures for ensembling. Two teams used two architectures each for two different tasks, for example one for semantic segmentation (binary classification between foreground and background pixels) and another for distance map prediction to separate touching nuclei. Notable innovations in model architectures tried by some of the top teams are described in Section IV-G.

### D. Model optimization

The choice of loss function depends on the desired output being predicted. Among various choices for the loss function, pixel-wise cross entropy was used by 28 teams for predicting binary or ternary probability maps, and it was by far the most popular loss function. Ten teams used Dice loss [49], and two teams used its variant such as smooth Jaccard index loss or IOU (intersection over union) loss [50]. For regression problems, seven teams used a smooth $L_1$ loss. Five teams used mean square error. In total, 16 teams used more than one loss function. Most teams trained their models end-to-end, except when an ensemble of more than one model was used, with the exception of team *Yunzhi* that used a cascade of two neural networks trained one after another.

### E. Post-processing

For post-processing, watershed segmentation (WS) was used by 17 teams. The most popular way to apply WS was on

the nuclear probability pixel map. Additionally, to separate touching nuclei several teams used a neural network to predict the location of a marker for each nucleus, such as by using a nuclear-core probability map, a distance map, or a vector map pointing to the nearest nuclear center. Cleaning up small or weakly detected nuclei was also a common theme. Non-maxima suppression and h-minima were commonly used along with a threshold to clean up false positives.

### *F. Training and testing time*

Training times ranged from 2 hours and 17 minutes on using a single Nvidia 1080Ti GPU for team *Junma* to 42 hours for team *Johannes Stegmaier* on a similar hardware. Testing times also had a wide range from 1 second per $1000 \times 1000$ image for team *Unblockabulls* on an Amazon Web Services GPU instance powered by an Nvidia K80 GPU to 2 minutes 58 seconds per image on an Nvidia Titan X GPU by team *CVBLab*.

### *G. Description of the top-five techniques*

We now describe the top-five techniques in more detail as examples of the innovations and diligence with which the participants tried to get robust generalization. Comparative results of the top-five techniques are shown in Figure 2. Specific details about parameter settings of each algorithm can be found in their respective manuscripts available on the challenge webpage [18] under the "manuscripts" tab.

*1) CUHK & IMSIGHT:* Extensive data augmentation based on random affine transform, rotation, and color jitter was used. Nuclei segmentation task was split into that of nucleus and boundary segmentation. A contour information aggregation network (CIA-Net), inspired by FCN [3] and U-net [2], to simultaneously segment nuclei and boundary was developed using Resnet50 [5] as the backbone architecture. The binary cross-entropy loss function that combined nucleus and boundary annotation errors was used to train the network. This algorithm missed some of the smaller nuclei and over-segmented (incorrectly splitting a large nuclei into multiple smaller nuclei) some larger nuclei as shown in Figure 2.

*2) BUPT.J.LI:* Images were color normalized and training data was augmented using random cropping, flipping, rotation, scaling, and noise addition. Deep layer aggregation [46] architecture was used to perform three tasks – (1) detect inside-nuclei pixels, (2) estimate the geometric center of the inside-nuclei pixels and (3) estimate a center vector that pointed towards the estimated nuclei center for each inside-nuclei pixel. During inference, the detected nuclei centers and center vectors were used to assign inside-nuclei pixels to one of the overlapping or touching nuclei instances. Since, nuclei boundary information was not explicitly used by the network, this technique produced overly smooth nuclei boundaries (Figure 2), especially for nuclei with high curvature boundaries.

*3) pku.hzq:* Extensive data augmentation was used such as flips, rotations, scaling, and noise addition. Then a U-Net [2] was used to predict a ternary class map similar to Kumar *et al*. [1]. Additionally, an MRCNN [4] was used for top-down instance segmentation. Predictions from the two models were combined as an ensemble for both boundary and nucleus prediction. Then the ensembled nuclei center masks were calculated using morphological eroding of the predicted nuclei pixels. A random walker was used to obtain instance segmentation masks from the ensembled semantic masks and center masks. From Figure 2, it is evident the boundaries for touching and overlapping nuclei were sometimes unnatural (and occasionally merged) due to pixel-level (semantic) ensembling of the boundary class predictions.

*4) Yunzhi:* For data preparation contrast-limited adaptive histogram equalization (CLAHE) [44] was used. Data augmentation was done using mirror flipping, rotations that were multiples of 90 degrees, color jitter, Gaussian noise addition, and elastic deformation. For each pixel, the probability of it belonging to a nucleus, or a nucleus boundary and unit vector to the center of the nuclei was computed using two cascaded U-nets [2]. First U-net predicted the inside nuclei pixels and unit vector to the center of the nuclei, which were then used in the subsequent U-net to accurately predict nuclei boundaries. Delineation of touching and overlapping nuclei using this technique heavily relied on accurate estimation of the unit vector that pointed towards the center of a nuclei and due to inaccuracy in precisely estimating the unit vector, this technique produced some over-segmentation and under-segmentation (incorrectly merging two touching or overlapping nuclei) errors (see Figure 2).

*5) Navid Alemi:* A neural network predicted both foreground (nuclear core) and background (nuclear boundary) markers. The neural network was a multi-scale feature-sharing network that used extensive skip connections, and was dubbed SpaghettiNet. For training the marker head prediction, the network used a combination of weighted Dice and binary cross entropy loss. For predicting the boundaries, it used smooth Jaccard loss and the boundary map was cleaned up using Frangi vesselness filter [51]. Finally, marker-controlled watershed segmentation using predicted markers and boundaries was employed to obtain the instance segmentation maps. Figure 2 shows that this technique produced overly smooth boundaries with some over-segmentation and under-segmentation errors.

### *H. Ensemble of top-five techniques*

Unlike ensembling of semantic segmentation, where class probabilities or decisions can be averaged for each pixel location, ensembling of instance segmentation results is far from trivial. Hence, we developed our own approach to generate the ensemble output of instances segmented by the top-five techniques because literature on this topic is thin and unconvincing. First, we looped over instances of the top-ranked technique and identified the corresponding nuclei instances from the other four techniques on the basis of maximum overlapping pixels. Once the matched instances from all techniques were identified, the corresponding ensemble instance was computed through pixel level majority voting as would be done for semantic segmentation of a single nucleus. Once we looped over all nuclei instances predicted by rank 1 technique, to incorporate the instances missed by rank 1 technique, we looped over all those instances of rank 2

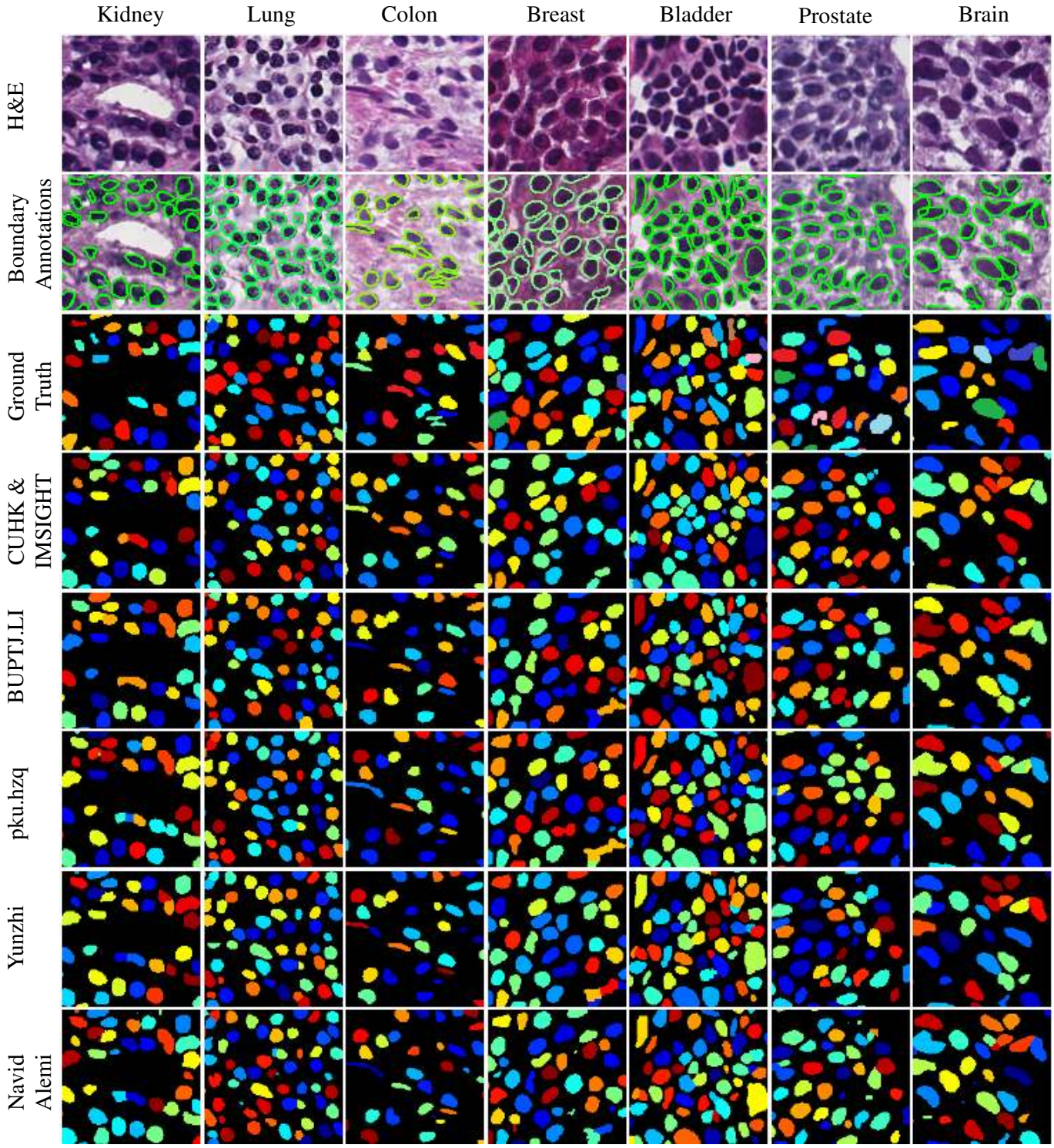


Fig. 2: Test sub-images taken from different organs exemplifying challenges of working with varied nuclear appearances and crowding patterns are shown in columns. Original H&E images, nuclear boundary annotations and segmentation results from the top-five techniques are shown in rows.

technique that did not find an overlap with those of the rank 1 technique. The process was repeated for rank 3 technique, but not for the other two remaining techniques because the extra instances detected by those two would not have a majority vote from the top three techniques. This ensembling method gave an overall a-AJI of 0.693 (95%CI: 0.682-0.703), which is only marginally better than the individual results of top-five teams.

### *I. Comparison to inter-human agreement*

We re-annotated all 14 test images and computed their a-AJI with the previous annotations. The re-annotation protocol was identical to the one used for creating the training set of MoNuSeg 2018 and the annotator was blinded to the previous test set annotations. The a-AJI between new and old manual annotations across 14 test images was 0.653 (95%CI 0.639-0.667), to which the a-AJI of the top few techniques compares very favorably. This suggests that for nucleus segmentation in H&E images, machine performance is at par with human performance if the image quality is as good as the one used in this challenge.

## V. Discussion and Conclusion

Some clear trends emerged from analyzing the top few techniques in Table III. While based on a prior idea that color normalization can improve performance of segmentation tasks [1], [52], it is becoming apparent that color augmentation (jitter) trains more robust segmentation models [53]. Most of the top techniques relied on heavy data augmentation including affine transformations, color jitter and noise addition. ResNet [5] seems to be an architecture of choice for several top performers irrespective of how they formulated the learning task. This is because the residual skip connections in ResNet allow backpropagation of gradient deep into the network without dilution. Most of the highly successful networks stuck to predicting pixel-wise class probabilities or using MRCNN [4] to predict instance maps. Watershed segmentation was among the most heavily utilized post-processing techniques. It was applied to the nuclear probability maps, most often coupled with a marker, where the marker was based on detecting the cores of individual nuclei. Some of the aforementioned general trends observed corroborated those found in instance segmentation challenges of general photography images such as Common Objects in Context (COCO) Challenge [35].

Although, the participating nuclei segmentation techniques reported significant improvement over the baseline method of [1], more improvements are possible and welcome. To further improve the nuclei segmentation quality, the ambiguity at the boundaries of touching and overlapping nuclei need to be better resolved. Additionally, new techniques should also produce more accurate nuclei boundaries without smoothing out high curvature boundaries. Another direction to be investigated is that of developing techniques that are tolerant of errors in the ground truth annotation itself. The role of generatative adverserial networks (GANs) to further improve nuclei segmentation performance should also be explored [31]. Based on the fact that the top techniques submitted to the MoNuSeg challenge had a-AJIs that were at par with that of a human annotator, it seems that it is time to put some of these techniques to use in nuclear morphometry based disease assessment studies to develop morphometric biomarkers. Finally, the robustness of the dataset and the techniques that have emerged as a part of the MoNuSeg challenge should be assessed for segmenting nuclei under multi-resolution and multi-stain settings. This can be achieved by conducting future competitions on the datasets containing annotated nuclei from images obtained at multiple microscopic resolutions (e.g, 10×, 20×, 40×, etc.) and including annotated nuclei from images stained with different types of stains (e.g. multiple IHC stains).

## Acknowledgements

This work was supported by the NCI-NIH (USA) grant no. 5R25-CA057699 for Kumar, NRF (Korea) grant no. 2016R1C1B2012433 for Vu and Kwak, and NIBIB-NIH (USA) grant no. R01EB020683 for Iftekharuddin. We are thankful to Gaurav Patel, Yashodhan Ghadge, and Sanjay Kumar for annotating nuclei for the testing set and to NVIDIA for donating the GPUs.

*N. Kumar is with the Department of Pathology, University of Illinois at Chicago.

R. Verma is with the Department of Biomedical Engineering, Case Western Reserve University.

D. Anand is with the Department of Electrical Engineering, Indian Institute of Technology Bombay.

Y. Zhou is with the Department of Computer Science and Engineering, Chinese University of Hong Kong.

O. F. Onder, E. Tsougenis and H. Chen are with Imsight Medical Technology Inc., Hong Kong.

P. A. Heng is with the Department of Computer Science and Engineering, Chinese University of Hong Kong.

J. Lui is with the School of Computer Science, Beijing University of Post and Telecommunications, Beijing, China.

Z. Hu is with the School of Electronics Engineering and Computer Science, Peking University, China.

Y. Wang is with the Department of Electrical and Computer Engineering, University of Oklahoma, Oklahoma, USA.

N. A. Koohbanani, M. Jahanifar, N. Z. Tajeddin, A. Gooya, S. Graham and N. Rajpoot are with the Department of Computer Science, University of Warwick, Warwick, United Kingdom.

X. Ren and Q. Wang are with the School of Biomedical Engineering, Shanghai Jiao Tong University, Shangai, China.

S. Zhou and D. Shen are with the Department of Radiology and BRIC, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, USA.

C. K. Yang, C. H. Weng, W. H. Yu and C. Y. Yeh are with aetherAI, Taipei City, Taiwan.

S. Yang is with Zhejiang University, Hangzhou, Zhejiang, China.

S. Xu is with Sun Yat-Sen University Cancer Center, Guangzhou, China and with Bio-totem Pte. Ltd, Shenzhen, China.

P. H. Yeung is with the University of Hong Kong, Hong Kong and with Bio-totem Pvt. Ltd, Shenzhen, China.

P. Sun is with Sun Yat-Sen University Cancer Center, Guangzhou, China.

A. Mahbod and I. Ellinger are with the Institute for Pathophysiology and Allergy Research, Medical University of Viena, Viena, Austria.

G. Schaefer is with the Department of Computer Science, Loughborough University, Loughborough, United Kingdom.

R. Ecker is with TissueGnostics GmbH, Vienna, Austria.

O. Smedby and C. Wang are with the Department of Biomedical Engineering and Health Systems, KTH Royal Institute of Technology, Stockholm, Sweden.

B. Chidester and J. Ma are with the School of Computer Science, Carnegie Mellon University, Pennsylvania, USA.

T. V. Ton and M.N. Do are with the Department of Electrical and Computer Engineering, University of Illinois at Urbana-Champaign, Illinois, USA.

M. T. Tran is with University of Science, Vietnam National University, Vietnam.

Q. D. Vu and J. T. Kwak are with the College of Software Convergence, Sejong University, Seoul, South Korea.

A. Gunda is with the Department of Mechanical Engineering, Indian Intitute of Technology Madras, Chennai, India.

R. Chunduri is with the Department of Aerospace Engineering, Indian Institute of Technology Bombay, Mumbai, India.

C. Hu. is with the Department of Computer Science, University of California Berkeley, California, USA.

X. Zhou. is with the Hong Kong University of Science and Technology, Hong Kong.

D. Lofti and R. Safdari are with Tehran Science and Research and Qazvin Branches of the Islamic Azad University, Iran.

A. Kascenas and A. O'Neil are with Canon Medical Research Europe, Edinburgh, United Kingdom.

D. Eschweiler and J. Stegmaier are with the Institute of Imaging and Computer Vision, RWTH Aachen University, Aachen, Germany.

Y. Cui and X. Tian are with the University of Science and Technology of China, Anhui, China.

B. Yin and K. Chen are with iFlytek AI Research, Guangzhou, China.

P. Gruening and E. Barth are with the Department of Computer Science, Institute for Neuro- and Bioinformatics, University of Lübeck, Lübeck, Germany.

E. Arbel, I. Remer and A. Ben-Dor are with Agilent Labs, Agilent Technologies Ltd., Tel-Aviv, Israel.

E. Sirazitdinova, M. Kohl and S. Braunewell are with Konica Minolta Laboratory Europe, Munich, Germany.

Y. Li, X. Xie and L. Shen are with the Computer Vision Institute, Shenzhen University, Shenzhen, China.

J. Ma is with the Department of Mathematics, Nanjing University of Science and Technology, Nanjing, China.

K. D. Baksi is with Biosciences R&D, TCS Research, TATA Consultancy Services Ltd., Pune, India.

M. A. Khan and J. Choo are with the Department of Computer Science and Engineering, Korea University, Seoul, South Korea.

A. Colomer and V. Naranjo are with Instituto de Investigación e Innovación en Bioingeníera, Universitat Politècnica de València, València, Spain.

L. Pei. and K. M. Iftekharuddin are with the Department of Electrical & Computer Engineering, Old Dominion University, Norfolk, Virginia.

K. Roy, and D. Bhattacharjee are with the Department of Computer Science and Engineering, Jadavpur University, Kolkata, India.

A. Pedraza and M. G. Bueno are with the Visilab Research Group, University of Castilla - La Mancha, Ciudad Real, Spain.

S. Devanathan, S. Radhakrishnan and P. Koduganty are with Cognizant Technology Solutions India Private Ltd, India.

Z. Wu is with Xiamen University, Xiamen, China.

G. Cai, X. Liu and Y. Wang are with the Department of Computer Science, Tongji University, Shanghai, China.

A. Sethi is with the Department of Electrical Engineering, Indian Institute of Technology Bombay, Mumbai, India.

## REFERENCES

[1] N. Kumar, R. Verma, S. Sharma, S. Bhargava, A. Vahadane, and A. Sethi, "A dataset and a technique for generalized nuclear segmentation for computational pathology," *IEEE Transactions on Medical Imaging*, vol. 36, no. 7, pp. 1550–1560, July 2017.

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

[3] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440.

[4] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2961–2969.

[5] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

[6] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[7] H. Chang, J. Han, A. Borowsky, L. Loss, J. W. Gray, P. T. Spellman, and B. Parvin, "Invariant delineation of nuclear architecture in glioblastoma multiforme for clinical and molecular association," *IEEE Transactions on Medical Imaging*, vol. 32, no. 4, pp. 670–682, April 2013.

[8] P. Filipczuk, T. Fevens, A. Krzyak, and R. Monczak, "Computer-aided breast cancer diagnosis based on the analysis of cytological images of fine needle biopsies," *IEEE Transactions on Medical Imaging*, vol. 32, no. 12, pp. 2169–2178, Dec 2013.

[9] A. H. Beck, A. R. Sangoi, S. Leung, R. J. Marinelli, T. O. Nielsen, M. J. van de Vijver, R. B. West, M. van de Rijn, and D. Koller, "Systematic analysis of breast cancer morphology uncovers stromal features associated with survival," *Science Translational Medicine*, vol. 3, no. 108, pp. 108ra113–108ra113, 2011. [Online]. Available: http://stm.sciencemag.org/content/3/108/108ra113

[10] A. Sethi, L. Sha, N. Kumar, V. Macias, R. J. Deaton, and P. H. Gann, "Computer vision detects subtle histological effects of dutasteride on benign prostate," *BJU international*, vol. 122, no. 1, pp. 143–151, 2018.

[11] J. H. Xue and D. M. Titterington, "t -tests, f -tests and otsu's methods for image thresholding," *IEEE Transactions on Image Processing*, vol. 20, no. 8, pp. 2392–2396, Aug 2011.

[12] X. Yang, H. Li, and X. Zhou, "Nuclei segmentation using marker-controlled watershed, tracking using mean-shift, and kalman filter in time-lapse microscopy," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 53, no. 11, pp. 2405–2414, Nov 2006.

[13] M. Veta, P. J. van Diest, R. Kornegoor, A. Huisman, M. A. Viergever, and J. P. Pluim, "Automatic nuclei segmentation in h&e stained breast cancer histopathology images," *PloS one*, vol. 8, no. 7, p. e70221, 2013.

[14] A. Vahadane and A. Sethi, "Towards generalized nuclear segmentation in histological images," in *Bioinformatics and Bioengineering (BIBE), 2013 IEEE 13th International Conference on*, Nov 2013, pp. 1–4.

[15] J. Schindelin, I. Arganda-Carreras, E. Frise, V. Kaynig, M. Longair, T. Pietzsch, S. Preibisch, C. Rueden, S. Saalfeld, B. Schmid, J.-Y. Tinevez, D. J. White, V. Hartenstein, K. Eliceiri, P. Tomancak, and A. Cardona, "Fiji: an open-source platform for biological-image analysis," *Nature Methods*, vol. 9, no. 7, pp. 676–682, Jul. 2012. [Online]. Available: http://dx.doi.org/10.1038/nmeth.2019

[16] A. E. Carpenter, T. R. Jones, M. R. Lamprecht, C. Clarke, I. H. Kang, O. Friman, D. A. Guertin, J. H. Chang, R. A. Lindquist, J. Moffat, P. Golland, and D. M. Sabatini, "Cellprofiler: image analysis software for identifying and quantifying cell phenotypes," *Genome Biology*, vol. 7, no. 10, p. R100, 2006. [Online]. Available: http://dx.doi.org/10.1186/gb-2006-7-10-r100

[17] "The cancer genome atlas (tcga)," http://cancergenome.nih.gov/.

[18] "Multi-organ nuclei segmentation challenge (MoNuSeg) 2018 [online]," https://monuseg.grand-challenge.org/, accessed 11 Feb. 2019.

[19] R. Verma, N. Kumar, A. Sethi, and P. H. Gann, "Detecting multiple subtypes of breast cancer in a single patient," in *2016 IEEE International Conference on Image Processing (ICIP)*, Sept 2016, pp. 2648–2652.

[20] N. Kumar, R. Verma, A. Arora, A. Kumar, S. Gupta, A. Sethi, and P. H. Gann, "Convolutional neural networks for prostate cancer recurrence prediction," in *Medical Imaging 2017: Digital Pathology*, vol. 10140. International Society for Optics and Photonics, 2017, p. 101400H.

[21] S. Ali and A. Madabhushi, "An integrated region-, boundary-, shape-based active contour for multiple object overlap resolution in histological imagery," *IEEE Transactions on Medical Imaging*, vol. 31, no. 7, pp. 1448–1460, July 2012.

[22] Y. Al-Kofahi, W. Lassoued, W. Lee, and B. Roysam, "Improved automatic detection and segmentation of cell nuclei in histopathology images," *IEEE Transactions on Biomedical Engineering*, vol. 57, no. 4, pp. 841–852, April 2010.

[23] S. Wienert, D. Heim, K. Saeger, A. Stenzinger, M. Beil, P. Hufnagl, M. Dietel, C. Denkert, and F. Klauschen, "Detection and segmentation of cell nuclei in virtual microscopy images: a minimum-model approach," *Scientific reports*, vol. 2, p. 503, Nov. 2012.

[24] H. Kong, M. Gurcan, and K. Belkacem-Boussaid, "Partitioning histopathological images: An integrated framework for supervised color-texture segmentation and cell splitting," *IEEE Transactions on Medical Imaging*, vol. 30, no. 9, pp. 1661–1677, Sept 2011.

[25] H. Chang, J. Han, A. Borowsky, L. Loss, J. W. Gray, P. T. Spellman, and B. Parvin, "Invariant delineation of nuclear architecture in glioblastoma multiforme for clinical and molecular association," *IEEE Transactions on Medical Imaging*, vol. 32, no. 4, pp. 670–682, April 2013.

[26] M. Zhang, T. Wu, and K. M. Bennett, "Small blob identification in medical images using regional features from optimum scale," *IEEE Transactions on Biomedical Engineering*, vol. 62, no. 4, pp. 1051–1062, April 2015.

[27] Y. Song, L. Zhang, S. Chen, D. Ni, B. Lei, and T. Wang, "Accurate segmentation of cervical cytoplasm and nuclei based on multiscale convolutional network and graph partitioning," *IEEE Transactions on Biomedical Engineering*, vol. 62, no. 10, pp. 2421–2433, Oct 2015.

[28] F. Xing, Y. Xie, and L. Yang, "An automatic learning-based framework for robust nucleus segmentation," *IEEE Transactions on Medical Imaging*, vol. 35, no. 2, pp. 550–566, Feb 2016.

[29] K. Sirinukunwattana and S. E. A. R. et. *al.*, "Locality sensitive deep learning for detection and classification of nuclei in routine colon cancer histology images," *IEEE Transactions on Medical Imaging*, vol. 35, no. 5, pp. 1196–1206, May 2016.

[30] P. Naylor, M. La, F. Reyal, and T. Walter, "Segmentation of nuclei in histopathology images by deep regression of the distance map," *IEEE Transactions on Medical Imaging*, vol. 38, no. 2, pp. 448–459, Feb 2019.

[31] F. Mahmood, D. Borders, R. Chen, G. N. McKay, K. J. Salimian, A. Baras, and N. J. Durr, "Deep adversarial training for multi-organ nuclei segmentation in histopathology images," *arXiv preprint arXiv:1810.00236*, 2018.

[32] H. Irshad, A. Veillard, L. Roux, and D. Racoceanu, "Methods for nuclei detection, segmentation, and classification in digital histopathology: A review, current status and future potential," *IEEE Reviews in Biomedical Engineering*, vol. 7, pp. 97–114, 2014.

[33] F. Xing and L. Yang, "Robust nucleus/cell detection and segmentation in digital pathology and microscopy images: A comprehensive review," *IEEE Reviews in Biomedical Engineering*, vol. 9, pp. 234–263, 2016.

[34] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet Large Scale Visual Recognition Challenge," *International Journal of Computer Vision (IJCV)*, vol. 115, no. 3, pp. 211–252, 2015.

[35] T.-Y. Lin and M. e. e. Maire, "Microsoft coco: Common objects in context," in *European conference on computer vision*. Springer, 2014, pp. 740–755.

[36] G. Litjens, P. Bandi, B. Ehteshami Bejnordi, O. Geessink, M. Balkenhol, P. Bult, A. Halilovic, M. Hermsen, R. van de Loo, R. Vogels *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.

[37] J. Irvin, P. Rajpurkar, M. Ko, Y. Yu, S. Ciurea-Ilcus, C. Chute, H. Marklund, B. Haghgoo, R. Ball, K. Shpanskaya *et al.*, "Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison," *arXiv preprint arXiv:1901.07031*, 2019.

[38] A. Janowczyk and A. Madabhushi, "Deep learning for digital pathology image analysis: A comprehensive tutorial with selected use cases," *Journal of Pathology Informatics*, vol. 7, July 2016.

[39] H. Irshad, L. Montaser-Kouhsari, G. Waltz, O. Bucur, J. Nowak, F. Dong, N. W. Knoblauch, and A. H. Beck, "Crowdsourcing image annotation for nucleus detection and segmentation in computational pathology: evaluating experts, automated methods, and the crowd." in *Pacific Symposium on Biocomputing*, 2015, pp. 294–305.

[40] E. D. Gelasca, B. Obara, D. Fedorov, K. Kvilekval, and B. Manjunath, "A biosegmentation benchmark for evaluation of bioimage analysis methods," *BMC Bioinformatics*, vol. 10, no. 1, p. 368, 2009.

[41] A. Vahadane, T. Peng, A. Sethi, S. Albarqouni, L. Wang, M. Baust, K. Steiger, A. M. Schlitter, I. Esposito, and N. Navab, "Structure-preserving color normalization and sparse stain separation for histological images," *IEEE Transactions on Medical Imaging*, vol. 35, no. 8, pp. 1962–1971, Aug 2016.

[42] M. Macenko, M. Niethammer, J. S. Marron, D. Borland, J. T. Woosley, X. Guan, C. Schmitt, and N. E. Thomas, "A method for normalizing histology slides for quantitative analysis," in *2009 IEEE International Symposium on Biomedical Imaging: From Nano to Macro*. IEEE, 2009, pp. 1107–1110.

[43] E. Reinhard, M. Adhikhmin, B. Gooch, and P. Shirley, "Color transfer between images," *IEEE Computer graphics and applications*, vol. 21, no. 5, pp. 34–41, 2001.

[44] S. M. Pizer, E. P. Amburn, J. D. Austin, R. Cromartie, A. Geselowitz, T. Greer, B. ter Haar Romeny, J. B. Zimmerman, and K. Zuiderveld, "Adaptive histogram equalization and its variations," *Computer vision, graphics, and image processing*, vol. 39, no. 3, pp. 355–368, 1987.

[45] G. Huang, Z. Liu, and K. Q. Weinberger, "Densely connected convolutional networks," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2261–2269, 2017.

[46] F. Yu, D. Wang, E. Shelhamer, and T. Darrell, "Deep layer aggregation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 2403–2412.

[47] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia, "Path aggregation network for instance segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8759–8768.

[48] V. Iglovikov and A. Shvets, "Ternausnet: U-net with vgg11 encoder pre-trained on imagenet for image segmentation," *arXiv preprint arXiv:1801.05746*, 2018.

[49] L. R. Dice, "Measures of the amount of ecologic association between species," *Ecology*, vol. 26, no. 3, pp. 297–302, 1945.

[50] P. Jaccard, "Étude comparative de la distribution florale dans une portion des alpes et des jura," *Bull Soc Vaudoise Sci Nat*, vol. 37, pp. 547–579, 1901.

[51] A. F. Frangi, W. J. Niessen, K. L. Vincken, and M. A. Viergever, "Multiscale vessel enhancement filtering," in *International conference on medical image computing and computer-assisted intervention*. Springer, 1998, pp. 130–137.

[52] A. S. et. *al.*, "Empirical comparison of color normalization methods for epithelial-stromal classification in h and e images," *Journal of Pathology Informatics*, vol. 7, 2016.

[53] D. Tellez, G. Litjens, P. Bandi, W. Bulten, J.-M. Bokhorst, F. Ciompi, and J. van der Laak, "Quantifying the effects of data augmentation and stain color normalization in convolutional neural networks for computational pathology," *arXiv preprint arXiv:1902.06543*, 2019.

[54] A. Janowczyk. (2018) On stain normalization in deep learning. [Online]. Available: http://www.andrewjanowczyk.com/on-stain-normalization-in-deep-learning/

TABLE III: Comparison of techniques that completed the MoNuSeg challenge

| Team Name | a-AJI (95% CI) | Pre-Proc. | | | | Data Augmentation | | | | | | | | | Model and Arch. | | | | | | | Loss | | | | Post-Proc. | | | Additional Notes |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Color Norm. | Unit Var. | Range Stand. | Hist. Eq. | Rotation | Flipping | Affine deform. | Scaling | Elastic deform. | Noise addition | Color jitter | Intensity jitter | Blur/sharpen | U-Net | Mask RCNN | FCN | PANet | ResNet | VGG-Net | DenseNet | Distance Map | Cross Entropy | Dice loss | L1 loss | L2 loss | Watershed seg. | Non-max supp. | Morph. ops. | |
| CUHK & IMSIGHT | 0.691 (0.680-0.702) | ✓ | | | | ✓ | | ✓ | | | | ✓ | | | ✓ | | ✓ | | | | | | ✓ | ✓ | | | | | | Macenko CN [42] |
| BUPT.J.LI | 0.687 (0.676-0.697) | ✓ | | | | ✓ | | ✓ | | | ✓ | | | | | | | | | | | | ✓ | | | ✓ | | | | SPCN [41]; Deep Layer Aggregation; geometric instance vector |
| pku.hzq | 0.685 (0.675-0.695) | ✓ | ✓ | | | ✓ | | ✓ | | | ✓ | | | | ✓ | ✓ | | | ✓ | | | | ✓ | | ✓ | | | ✓ | ✓ | Macenko CN [42]; ResNet with feature pyramid network |
| Yunzhi | 0.679 (0.668-0.690) | | ✓ | | ✓ | ✓ | ✓ | ✓ | | ✓ | ✓ | ✓ | | | ✓ | | | | ✓ | | | | ✓ | | | | ✓ | | ✓ | Cascaded U-Net with ResNet-like arch. |
| Navid Alemi | 0.678 (0.666-0.689) | | | | | ✓ | ✓ | ✓ | | ✓ | ✓ | ✓ | | ✓ | | | | | | | | | ✓ | ✓ | | | ✓ | | | Concatenated HSV and L channels to RGB multi-headed sphagetti net; smooth Jaccard index for boundary detection; boundary map cleaned by frangi vesselness filter gave markers |
| xuhuaren | 0.664 (0.652-0.676) | | ✓ | | | ✓ | | ✓ | | | ✓ | ✓ | | | | ✓ | | ✓ | | | | | ✓ | | ✓ | | | ✓ | | |
| aetherAI | 0.663 (0.653-0.673) | | | | | ✓ | ✓ | ✓ | ✓ | | ✓ | ✓ | | | | ✓ | | ✓ | ✓ | | ✓ | | ✓ | | ✓ | | | ✓ | | |
| Shuang Yang | 0.662 (0.652-0.673) | | | | | ✓ | | ✓ | | | ✓ | ✓ | | | | ✓ | | ✓ | ✓ | | | | ✓ | ✓ | | | | ✓ | ✓ | |
| Bio-totem & SYSUCC | 0.662 (0.652-0.672) | | | | ✓ | ✓ | | ✓ | | | ✓ | ✓ | | | | ✓ | | | ✓ | | | | ✓ | | ✓ | | | | | Color histogram equalization [54] |
| Amirreza Mahbod | 0.657 (0.649-0.666) | ✓ | | ✓ | | ✓ | | ✓ | | | ✓ | ✓ | | | ✓ | | | | | | | ✓ | ✓ | ✓ | | | ✓ | ✓ | ✓ | Macenko CN [42]; markers are filtered distance maps |
| CMU-UIUC | 0.656 (0.645-0.667) | ✓ | | | | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | | ✓ | ✓ | ✓ | | | | ✓ | | | | ✓ | ✓ | | | | | ✓ | SPCN [41]; in-place augmentation of segmented nuclei |
| Graham&Vu | 0.653 (0.643-0.663) | ✓ | | | | ✓ | | ✓ | | | ✓ | ✓ | ✓ | | | ✓ | | | ✓ | | ✓ | ✓ | ✓ | | ✓ | ✓ | | | | SPCN [41]; combined detection and distance map pred. |
| Unblockabulls | 0.651 (0.637-0.666) | ✓ | | | | ✓ | ✓ | | ✓ | | | | | | ✓ | | | | | | | | ✓ | | | | ✓ | | | Macenko CN [42], concatenated hematoxylin channel |
| Tencent AI Lab | 0.646 (0.635-0.657) | ✓ | | | | ✓ | ✓ | | ✓ | | ✓ | | ✓ | ✓ | ✓ | ✓ | | | ✓ | | | | ✓ | | ✓ | | | | | Edge enhancement used in pre-processing to separate nuclei |
| DeepMD | 0.633 (0.619-0.647) | ✓ | | | | ✓ | ✓ | | ✓ | | ✓ | | | ✓ | ✓ | | | | | ✓ | | | ✓ | ✓ | | | ✓ | | | SPCN [41]; random sharpening; TernausNet architecture |
| Canon Medical Research Europe | 0.633 (0.604-0.661) | | | ✓ | | ✓ | ✓ | | ✓ | ✓ | | | | | ✓ | | | | | ✓ | ✓ | | ✓ | | | | | | | |
| Johannes Stegmaier | 0.623 (0.603-0.643) | | | ✓ | | ✓ | ✓ | ✓ | ✓ | | | ✓ | ✓ | | ✓ | | | | | | | ✓ | ✓ | | | | ✓ | | | |
| Yanping | 0.623 (0.610-0.636) | ✓ | | | | ✓ | ✓ | | ✓ | | ✓ | | | ✓ | ✓ | ✓ | | | ✓ | | | | | | | | ✓ | | ✓ | SPCN [41] |
| Philipp Gruening | 0.621 (0.606-0.636) | | | ✓ | ✓ | ✓ | ✓ | | | | ✓ | | | | ✓ | | | | | ✓ | | | ✓ | | | | ✓ | | | Cross entropy and squared cosine losses |
| Agilent Labs | 0.618 (0.598-0.638) | | | | | ✓ | | ✓ | | ✓ | | ✓ | | | ✓ | | | | | | | ✓ | | | | ✓ | ✓ | ✓ | ✓ | |
| Konica Minolta Lab EU | 0.611 (0.601-0.622) | | | | | ✓ | ✓ | | | ✓ | | | | | ✓ | | | | | ✓ | | | | ✓ | | | | | ✓ | |
| OnePiece | 0.606 (0.592-0.620) | ✓ | | | | | | | | | | | | | ✓ | | | | | | | | ✓ | | | | ✓ | | | SPCN [41] |
| Junma | 0.593 (0.581-0.606) | ✓ | | ✓ | | | ✓ | | | | | | | | | | | | | | | | | ✓ | | | ✓ | | | SPCN [41], blue channel extraction |
| Biosciences R&D, TCS | 0.578 (0.538-0.619) | | | | | ✓ | ✓ | | | | | | | | ✓ | | | | | | | | | | | | ✓ | | | Smooth Jaccard index; separate nuclei using convexity |
| Azam Khan | 0.575 (0.556-0.594) | | | | | ✓ | ✓ | ✓ | | | | ✓ | | ✓ | | ✓ | | ✓ | | | | | ✓ | ✓ | | | | ✓ | | |
| CVBLab | 0.574 (0.560-0.588) | ✓ | | | | | | | | | | | | | ✓ | | | | | | | | | ✓ | | | ✓ | | | SPCN [41] |
| Linmin Pei | 0.562 (0.548-0.577) | ✓ | | | | ✓ | ✓ | | | | | | | | ✓ | | | | | | | | | | | ✓ | ✓ | | ✓ | SPCN [41] |
| DB-KR-JU | 0.455 (0.428-0.481) | ✓ | | | | ✓ | | | | | | | | | | | | | | | | | ✓ | | | | ✓ | | ✓ | Reinhard [43]+Macenko [42] CN; separate nuc. using circularity |
| VISILAB | 0.444 (0.425-0.463) | ✓ | | | | ✓ | ✓ | | | | | | | | | | | | | ✓ | | | ✓ | | | | ✓ | | | Macenko CN [42] |
| Sabarinathan | 0.444 (0.424-0.464) | | | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | | | | ✓ | ✓ | ✓ | | | | | | | | ✓ | ✓ | | | | | | CLAHE [44] |
| Silvers | 0.278 (0.228-0.328) | | | | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | | ✓ | ✓ | | ✓ | | | | | | ✓ | | ✓ | | | | ✓ | | ✓ | Combination of DenseNet and U-Net |
| TJ | 0.130 (0.106-0.154) | ✓ | | | | ✓ | ✓ | | | ✓ | | | | | ✓ | ✓ | | | | ✓ | | | ✓ | | | | | | ✓ | SPCN [41]; ensemble of multiple architectures |