ARTICLES
https://doi.org/10.1038/s42256-019-0018-3
nature machine intelligence

An integrated iterative annotation technique for easing neural network training in medical image analysis

Brendon Lutnick[1], Brandon Ginley[1], Darshana Govind[1], Sean D. McGarry[2], Peter S. LaViolette[3], Rabi Yacoub[4], Sanjay Jain[5], John E. Tomaszewski[1], Kuang-Yu Jen[6] and Pinaki Sarder[1]*

Neural networks promise to bring robust, quantitative analysis to medical fields. However, their adoption is limited by the technicalities of training these networks and the required volume and quality of human-generated annotations. To address this gap in the field of pathology, we have created an intuitive interface for data annotation and the display of neural network predictions within a commonly used digital pathology whole-slide viewer. This strategy used a 'human-in-the-loop' to reduce the annotation burden. We demonstrate that segmentation of human and mouse renal micro compartments is repeatedly improved when humans interact with automatically generated annotations throughout the training process. Finally, to show the adaptability of this technique to other medical imaging fields, we demonstrate its ability to iteratively segment human prostate glands from radiology imaging data.

In the current era of artificial intelligence, robust automated image analysis is attained using supervised machine-learning algorithms. This approach has been gaining considerable ground in virtually every domain of data analysis, mainly since the advent of neural networks[1–4]. Neural networks are a broad range of graphical models, whose nodes are variably activated by a nonlinear operation on the sum of their inputs[3,5]. The connections between nodes are modulated by weights, which are adjusted to alter the contribution of that node to the network output. These weights are iteratively tuned via backpropagation so that the input of data leads to a desired output (usually a classification of the data)[6]. Particularly useful for image analysis are convolutional neural networks (CNNs)[2,3], a specialized subset of neural networks. CNNs leverage convolutional filters to learn spatially invariant representations of image regions specific to the desired image classification. This allows high-dimensional filtering operations to be learned automatically, a task that has traditionally been performed through hand-engineering. The potential of neural networks exceeds that of other machine-learning techniques[7], but they are problematic in certain applications. Namely, they require significant amounts of annotated data to provide generalized high performance.

Easing the burden of data annotation is arguably as important as generating state-of-the-art network architectures, which without sufficient data are unusable[8,9]. Many large-scale modern machine-learning applications are based on cleverly designed crowd-sourced active-learning pipelines. In an era of constant firmware updates, this advancement comes in the form of human-in-the-loop training[10–12]. Initiated by low classification probabilities, machine-learning applications, such as automated teller machine character recognition, self-driving cars and Facebook's automatic tagging, all rely on user-refined training sets for fine-tuning neural network applications post deployment[3]. These 'active learning' techniques require users to 'correct' the predictions of a network, identifying gaps in network performance[13].

Although computational strategies for image analysis are increasingly being translated to biological research, the application of neural networks to biological datasets has lagged their implementation in computer science[14,15]. This late adoption of CNN-based methods is largely due to the lack of centrally curated and annotated biological training sets[16]. Due to the specialized nature of medical datasets, the expert annotation needed to generate training sets is less feasible than for traditional datasets[17]. This issue creates challenges when trying to apply CNNs to medical imaging databases, where domain-expert knowledge is required to perform image annotation. This annotation is expensive, time-consuming and labour-intensive, and there are no technical media that enable easy transference of this information from clinical practice to training sets[18].

Despite the challenges, using neural networks to segment and classify tissue slides can aid clinical diagnosis and help create improved diagnostic guidelines based on quantitative computational metrics. Moreover, neural networks can generate searchable data repositories[19], providing practicing clinicians and students access to previously unavailable collections of domain knowledge[20–22], such as labelled images and associated clinical outcomes. Achieving such access on a large scale will require a combination of curated pathological datasets, machine-learning classifiers[3], automatic anomaly detection[23,24] and efficiently searchable data hierarchies[21]. Finally, pipelines will be needed for creating easily viewable annotations on pathology images. Towards this aim, we have developed an iterative interface between the successful semantic segmentation network DeepLab v2[25] and the widely used whole-slide image (WSI) viewing software Aperio ImageScope[26], which we have termed Human AI Loop (H-AI-L) (Fig. 1). Put simply, the algorithm converts annotated regions stored in XML format (provided in ImageScope) into

[1]Department of Pathology & Anatomical Sciences, SUNY Buffalo, New York, NY, USA. [2]Department of Biophysics, Medical College of Wisconsin, Wauwatosa, WI, USA. [3]Department of Radiology and Biomedical Engineering, Medical College of Wisconsin, Wauwatosa, WI, USA. [4]Department of Medicine, Nephrology, SUNY Buffalo, New York, NY, USA. [5]Department of Medicine, Nephrology, Washington University School of Medicine, St Louis, MO, USA. [6]Department of Pathology, University of California, Davis Medical Center, Sacramento, CA, USA. *e-mail: pinakisa@buffalo.edu

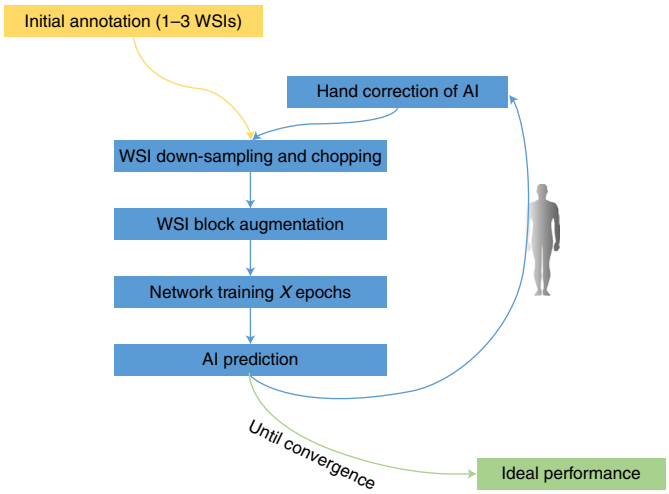

Fig. 1 | Iterative H-AI-L pipeline overview. Schematic representation of the H-AI-L pipeline for training semantic segmentation of WSIs. Several rounds of training are performed using human expert feedback to optimize ideal performance, resulting in improved efficiency in network training with limited numbers of initial annotated WSIs.

image region masks. These masks are used to train the semantic segmentation network, whose predictions are converted back to XML format for display in ImageScope. This graphical display of the network output is an ideal visualization tool for making segmentation predictions on WSIs. It allows the entire tissue slide to be viewed, with panning and zooming, and it uses the efficient JPG2000 decompression[27] of WSI files provided by ImageScope. Note that while the current code works only in ImageScope, the proposed system can easily be adapted for other WSI viewers, such as the universal viewer Pathcore Sedeen[28], as well as ImageJ. Note also that ImageScope and the DeepLab architecture are not currently approved for diagnostic procedures. Therefore, for any potential application of our system in a clinical workflow, our pipeline needs to be adopted using annotation and machine-learning tools that are currently approved for clinical diagnosis.

Using this open-sourced pipeline, a supervising domain expert can correct the network predictions (deleting false positives and annotating false-negative regions) before initiating further training using the newly annotated data. Thus, networks can be trained either 'on demand' or as the data become available. Using H-AI-L, we are able to significantly reduce the annotation effort required to learn robust segmentations of large microscopy images[28]. Adapting this technique to other modes of medical imaging is highly feasible, which we demonstrate using magnetic resonance imaging (MRI) data.

Results

To evaluate the utility of H-AI-L, we first quantified its performance and efficiency in segmenting histologic sections of kidney tissue, beginning with glomerular localization in mouse kidney WSIs[4,29–32]. This glomeruli segmentation network was trained for five iterations, using a combination of periodic acid–Schiff (PAS) and haematoxylin and eosin (H&E)-stained murine renal sections. For more data variation, streptozotocin (STZ)-induced diabetic nephropathy[33–36] murine data were included in iteration 4 (Table 1). To validate the performance of our network, we use four holdout WSIs, including one STZ-induced WSI.

During the training process, we observed approximately four- to tenfold increases in average glomerular annotation speed between the initial and end iterations (Fig. 2a). Compared to each annotator's baseline speed, these increases represent time savings of 81.4, 82 and 72.7% for annotators 1, 2 and 3, respectively. The prediction

Table 1 | H-AI-L segmentation mouse WSI training and testing datasets

H-AI-L dataset							
Annotation iteration		0	1	2	3	4	Test
WSIs added		1	2	4	6	4	4
Total glomeruli	Normal	32	84	86	418	0	138
	STZ	0	0	0	0	293	96

Mouse WSI training set used to train the glomerular segmentation network. Data presenting structural damage from STZ-induced diabetes[1] were introduced in iteration 4. The test dataset included three normal and one STZ-induced murine renal WSI.

performance increase is shown in Fig. 2b, where the network reaches nearly perfect performance on a holdout dataset by annotation iteration 4. One side effect of using iterative annotation is intuitive qualification of network performance after each interaction. That is, an expert interacts with the network predictions after each training round, visualizing network biases and shortcomings on holdout data. Two examples of evolving network predictions are highlighted in Supplementary Video 1.

To improve network prediction efficiency, we designed a two-stage segmentation approach. This uses two segmentation networks, first identifying hotspot regions at 1/16th scale and then segmenting them at the highest resolution. This approach (which we call multi-pass segmentation) provides a better F-measure (F1 score)[37,38] (Fig. 2b) than a full-resolution pass, as well as approximately 4.5-times faster predictions (Fig. 2c). An overview of this method can be found in Supplementary Fig. 1.

Quantification of the performance achieved by our method in WSIs is a challenge due to the imbalance between class distributions[39]. Therefore, we choose to report the F-measure, which considers both precision and recall (sensitivity) simultaneously[37], as specificity and accuracy are always high because the negative region is large with respect to the positive class. This choice of using the F-measure is particularly important considering the performance characteristics of multi-pass segmentation. During testing we found that the multi-pass approach trades segmentation sensitivity for increased precision, while outperforming full analysis overall, with an improved F1 score (Fig. 2). This result is due to a lower false-positive rate achieved by multi-pass segmentation as a result of the low-resolution network pre-pass, which limits the amount of background region seen by the high-resolution network. Overall (on four holdout WSIs), our network achieved its best performance after the fifth iteration of training using multi-pass segmentation, with a sensitivity of 0.92 ± 0.02, specificity of 0.99 ± 0.001, precision of 0.93 ± 0.14 and accuracy of 0.99 ± 0.001.

Network performance analysis is further complicated by human annotation errors. We note several instances where network predictions outperformed human annotators, despite being trained using flawed annotations. This phenomenon is highlighted in Fig. 3, where glomerular regions annotated manually in iteration 0 are compared to the iteration 5 network predictions. Such errors are more prevalent in WSIs annotated in early iterations, where network predictions need the most correction.

To qualitatively prove the effectiveness and extendibility of our method, we show its extension to multi-class detection by segmenting glomerular nuclei types[40,41] and interstitial fibrosis and tubular atrophy (IFTA)[42,43], as well as by differentiating sclerotic and non-sclerotic glomeruli[44]. This analysis is performed in mouse kidney and human renal biopsies. Figure 4 shows the glomeruli detection network from Fig. 2 adapted for nuclei detection. This study was carried out by retraining the high-resolution network using a set of 143 glomeruli with labelled podocyte and non-podocyte nuclei, marked via immunofluorescence labelling. For this analysis, the

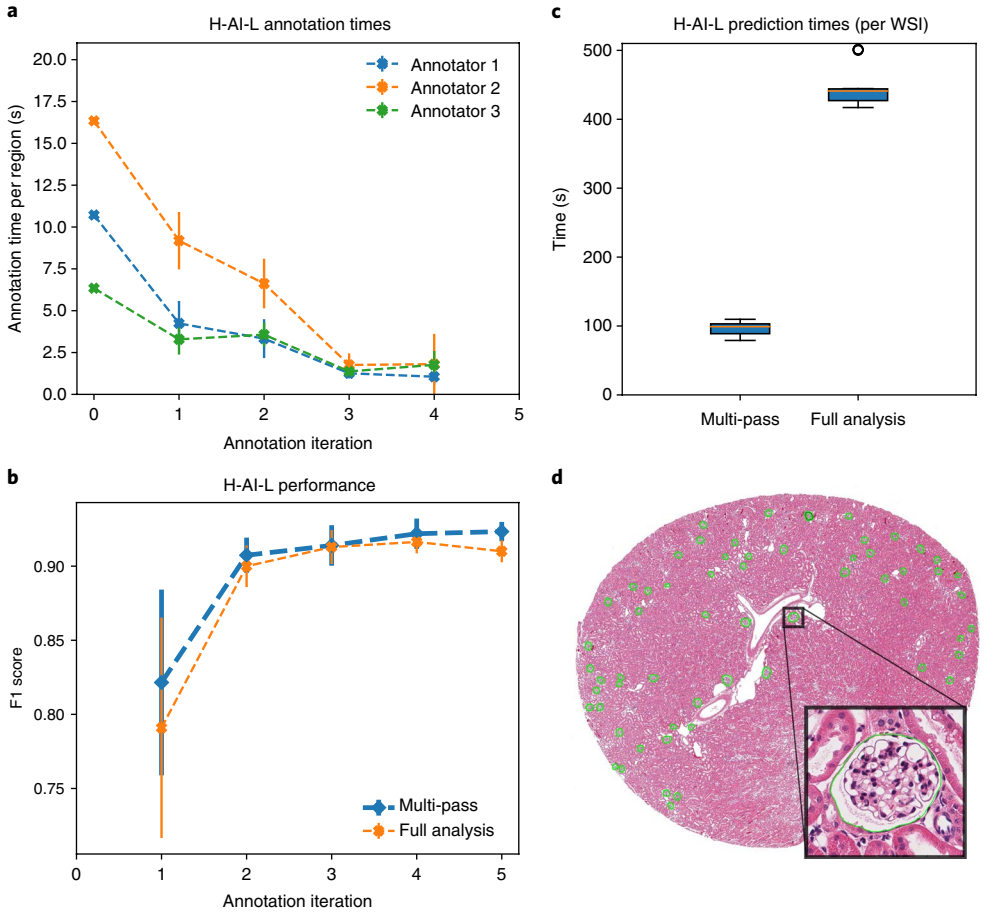

Fig. 2 | H-AI-L pipeline performance analysis for glomerular segmentation on holdout mouse WSIs. a, Average annotation time per glomerulus as a function of annotation iteration. The data are averaged per WSI and normalized by the number of glomeruli in each WSI. The 0th iteration was performed without pre-existing predicted annotations, whereas subsequent iterations use network predictions as an initial annotation prediction that can be corrected by the annotator. **b**, F1 score of glomerular segmentation of four holdout mouse renal WSIs as a function of training iteration. **c**, Run times for glomerular segmentation prediction on holdout mouse renal WSIs using H-AI-L with multi-pass (two-stage segmentation) versus full-resolution segmentation. **d**, Example of a mouse WSI with segmented glomeruli (×40, H&E-stained). Network predictions are outlined in green. The error bars indicate ±1 standard deviation.

low-resolution network from Fig. 2 was kept unchanged to identify the glomerular regions in the mouse WSI.

Due to the non-sparse nature of IFTA regions in some human WSIs, we forgo our multi-pass approach to generate the results shown in Fig. 5. The development of this IFTA network has been limited due to the biological expertise required to produce these multi-class annotations. However, preliminary segmentation results on holdout WSIs are promising, even though only 15 annotated biopsies were used for training (Fig. 5). We note that this is a small training set, as human biopsy WSIs contain much less tissue area than the mouse kidney sections used to train the glomerular segmentation network above.

Finally, to show the adaptability of the H-AI-L pipeline to other medical imaging modalities, we quantify the use of our approach for the segmentation of human prostate glands from T2 MRI data. These data were oriented and normalized as described in ref. [45] and saved as a series of TIFF image files. These images can be opened in ImageScope and are compatible with our H-AI-L pipeline. This analysis was completed using a training set of data from 39 patients, with an average of 32 slices per patient (512×512 pixels) (Fig. 6d); 509 of the total 1,235 slices contained prostate regions of interest. Iterative training was completed by adding data from four new patients to the training set before each iteration. Data from the remaining seven patients were used as a holdout testing set (a full breakdown is available in Supplementary Table 1). The newly annotated/corrected training data were augmented ten times, and a full-resolution network was trained for two epochs during each iteration: the results of this training are presented in Fig. 6. While the network performs well after just one round of training, the performance on holdout patient data continues to improve with the addition of training data (Fig. 6a), achieving a sensitivity of 0.88 ± 0.04, specificity of 0.99 ± 0.001, precision of 0.9 ± 0.03 and accuracy of 0.99 ± 0.001. This trend is also loosely reflected in the network prediction on newly added training data, where an upward trend in prediction performance is observed in Fig. 6b. Notably, when our iterative training pipeline is applied to this dataset, annotation is reduced by approximately 90% percent after the second iteration; only 10% of the MRI slices containing prostate fall below our segmentation performance threshold (Fig. 6c). We note that careful conversion between the DICOM and TIFF format (considering orientation and colour scaling) is essential for this analysis.

Conclusions

We have developed an intuitive pipeline for segmenting structures from WSIs commonly used in pathology, a field where there is often a large disconnect between domain experts and engineers.

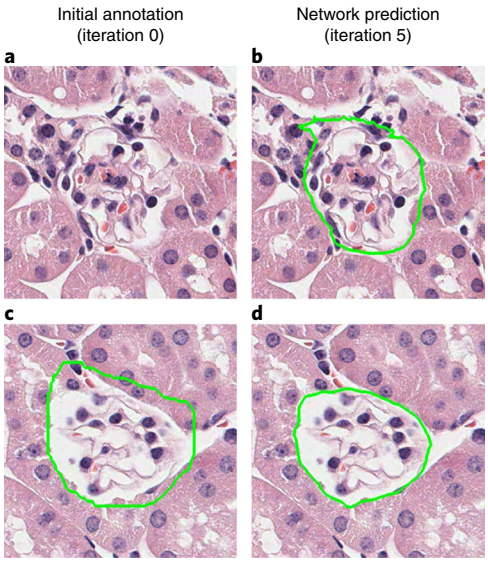

Fig. 3 | H-AI-L human annotation errors (mouse data). a–d, Comparison of initial manual annotations from iteration 0 (**a**,**c**) with their respective final network predictions from iteration 5 (**b**,**d**). These examples were selected due to poor manual annotation, where the glomerulus was not annotated (**a**) or showed poorly drawn boundaries (**c**). These images are captured at ×40, and tissue was stained using H&E.

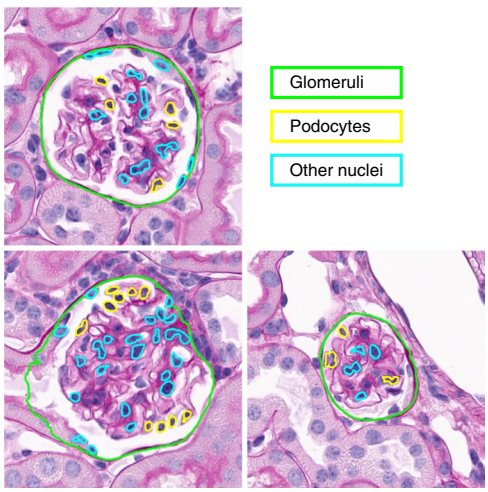

Fig. 4 | Multiclass nuclei prediction on a mouse WSI. Several examples of multi-class nuclei predictions are visualized on a mouse WSI (×40, PAS-stained). Here, transfer learning was used to adapt the high-resolution network from above (Fig. 2) to segment nuclei classes. This network was trained using 143 labelled mouse glomeruli. The low-resolution network was kept unchanged for the initial detection of glomeruli. We expect the results to significantly improve using more labelled training data.

To bridge this gap, we seek to provide pathologists with robust data analytics provided by state-of-the-art neural networks. We have developed an intuitive library for the adaptation of DeepLab v2[25], a semantic segmentation network, to WSI data commonly used in the field. This library uses annotation tools from the common WSI viewing software Aperio ImageScope[26] to annotate and display network predictions. Training, prediction and validation of the network are performed via a single Python script with a command line interface, making data management as simple as dropping data into a pre-determined folder structure.

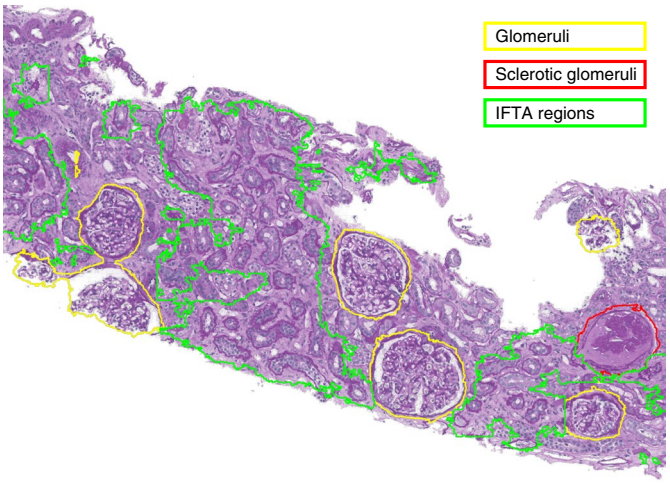

Fig. 5 | Multiclass IFTA prediction on a holdout human renal WSI. Segmentation of healthy and sclerotic glomeruli, as well as IFTA regions from human renal biopsy WSI (×40, PAS-stained). Due to the non-sparse nature of IFTA regions, these predictions were made using only a high-resolution pass. This is a screenshot of Aperio ImageScope, which we use to interactively visualize the network predictions.

Our iterative, human-in-the-loop training allows considerably faster annotation of new WSIs (or similar imaging data), because network predictions can easily be corrected in ImageScope before incorporation into the training set. With this approach, network performance can be qualitatively assessed after each iteration. Newly added data act as a holdout validation set, where predictions are easily viewed during correction. The theoretical performance achievable by this method is bounded by the training set used, and is therefore the same as the current state-of-the-art (manual annotation of all training data). However, due to the increased speed of annotation and the intuitive visualization of network performance (allowing selection of poorly predicted new data after each iteration), H-AI-L training can converge to the upper bound of performance more efficiently than the traditional method. That is, H-AI-L achieves state-of-the-art segmentation performance much faster than traditional methods, which are limited by data annotation speed (Fig. 7). Our H-AI-L approach offers an ideal viewing environment for network predictions on WSIs, using the fast pan and zoom functionality provided by ImageScope[27], improving the accuracy and ease of expert annotation.

The ability to transfer parameters from a trained network (repurposing it for a different task) ensures that segmentation of tissue structure can be tailored to any clinical or research definition, including other biomedical imaging modalities. Our two-stage segmentation (multi-pass) analysis allows rapid prediction of sparse regions from large WSIs, without sacrificing accuracy due to low-resolution analysis alone. Inspired by the way pathologists scan tissue slides, multi-pass approaches have been successfully described in digital pathology for detecting cell nuclei[46]. We believe that this technique offers the perfect compromise between speed and specificity, producing high-resolution sparse segmentations ideal for display in ImageScope. Our method provides non-sparse segmentation of WSIs by forgoing multi-pass analysis. However, in the future we plan to change how the class hierarchy is defined in our algorithm, offering easy functionality to search for low-resolution regions with high-resolution sub-compartments.

In the future, we will also extensively test our method in a clinical research setting. This testing will evaluate both the segmentation performance and ergonomic aspects affecting a clinician's ease of use. We will extend our method to provide anomaly detection,

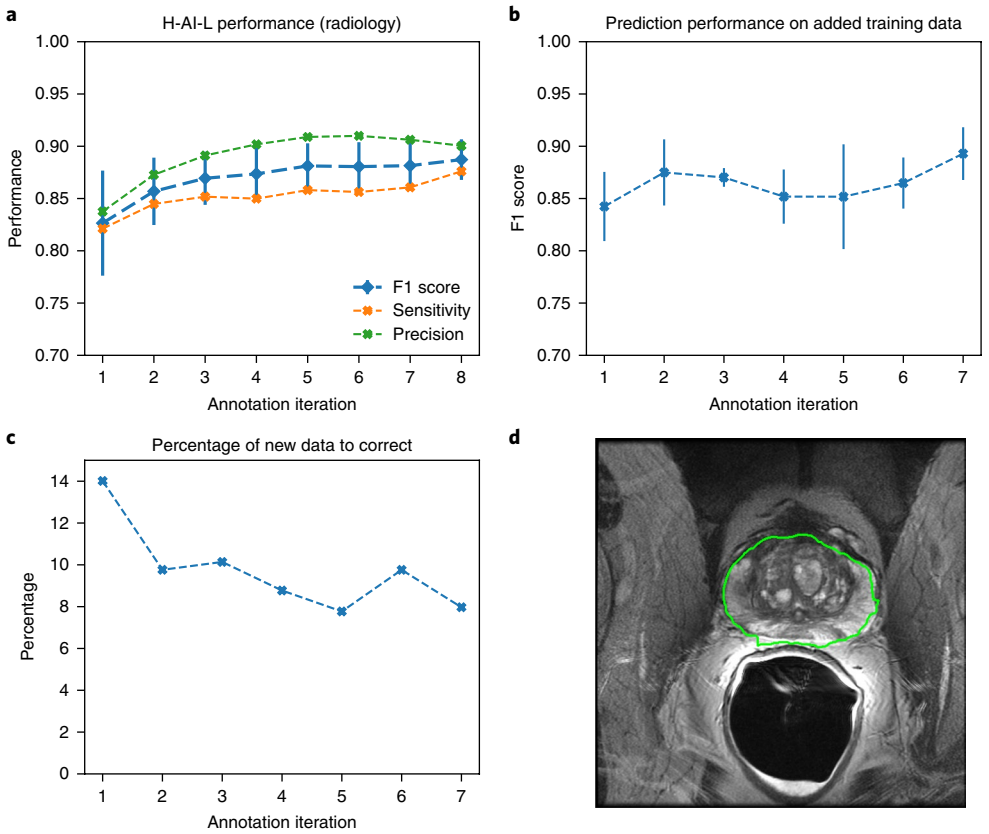

Fig. 6 | H-AI-L method performance analysis for human prostate segmentation from T2 MRI slices. a, Segmentation performance as a function of training iteration, evaluated on 7 patient holdout MRI images (224 slices). Performance was evaluated on a patient basis. We note that despite the decline in network precision after iteration 6, the F1 score improves as a result of increasing sensitivity. **b**, The prediction performance on added training data, before network training. This figure shows the prediction performance on newly added data with respect to the expert-corrected annotation, and is evaluated on a patient basis (data from four new patients were added at the beginning of each training iteration). **c**, The percentage of prostate regions where network prediction performance (F1 score) fell below an acceptable threshold (percentage of slices that needed expert correction) as a function of training iteration. We define acceptable performance as F1 score > 0.88. Using this criterion, expert annotation of new data is reduced by 92% by the fifth iteration. **d**, A randomly selected example of a T2 MRI slice with segmented prostate; the network predictions are outlined in green. The error bars indicate ±1 standard deviation. A detailed breakdown of the training and validation datasets is available in Supplementary Table 1.

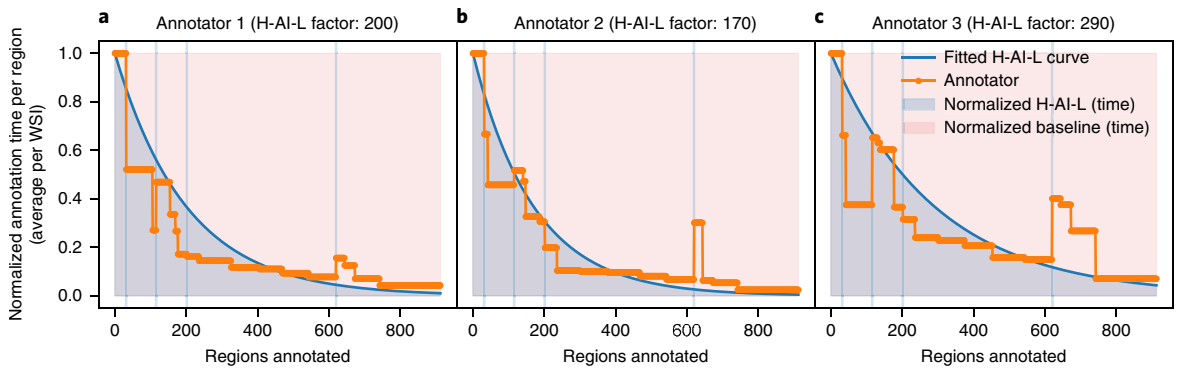

Fig. 7 | Annotation time-savings using the H-AI-L method while comparing to baseline segmentation speed. H-AI-L plots showing the annotation time per region normalized with respect to the baseline annotation speed of each annotator for the result shown in Fig. 2a. An exponential decay distribution (H-AI-L curve) is fitted to each annotator, where the H-AI-L factor is the exponential time constant: a derivation can be found in the Methods. The vertical lines are gaps between iterations (where the network was trained). The area under the H-AI-L curve represents the normalized annotation time per annotator. This can be compared to the area of the normalized baseline region, which represents the normalized annotation time without the H-AI-L method. **a**, The time-savings by annotator 1 (calculated to be 81.3%) when creating the training set used to train the glomerular segmentation network in Fig. 2. **b**, Annotator 2 was 82.0% faster. **c**, Annotator 3 was 72.7% faster. While the *y* axis in these plots is not a direct measure of network performance, it is highly correlated. The spike in annotation time seen at 600 regions is data from a WSI with severe glomerular damage from diabetic nephropathy. Future work will involve deriving optimal iterative training strategies based on information mined via such plots, with a goal of reducing annotation burdens for expert annotators.

defining a confidence metric and threshold where WSIs are flagged for further evaluation. Further, to minimize the expert's time, we will create an algorithm to predict the optimal amount of annotation performed in each iteration, using a curve fitting similar to Fig. 7. We will also adapt our method for native use with a DICOM viewer and a three-dimensional CNN for segmentation, allowing easier workflows for segmentation of radiology datasets, and mitigating the issues of data orientation and gamut mapping when converting to 8-bit TIFF images. Given these tools, we foresee a segmentation approach similar to our H-AI-L method underpinning efforts to build searchable medical image databases for research and education.

Methods

All animal tissue sections were collected in accordance with protocols approved by the Institutional Animal Care and Use Committee at the University at Buffalo, and in a manner consistent with federal guidelines and regulations and in accordance with recommendations of the American Veterinary Medical Association guidelines on euthanasia. Human renal biopsy samples were collected from the Kidney Translational Research Center at Washington University School of Medicine, directed by S.J., following a protocol approved by the Institutional Review Board at the University at Buffalo before commencement. Digital MRI images of human prostate glands were provided by P.S.L., following a protocol approved by the Institutional Review Board at the Medical College of Wisconsin. All human methods were performed in accordance with the relevant federal guidelines and regulations. All patients provided written informed consent.

For mouse pathology sample preparation, C57BL/6J background mice were euthanized, and their kidneys were perfused, extracted and embedded in paraffin. Mice were either treated with STZ to induce diabetic nephropathy or with an STZ vehicle for control. The murine WSIs used (Figs. 2 and 3) were sliced from paraffin-embedded kidney sections at 2 µm, stained with either PAS or H&E, and bright-field imaged at 0.25 µm per pixel resolution and ×40 magnification using a whole-slide scanner (Aperio Scan Scope, Leica). The sections used for podocyte segmentation (Fig. 4) were prepared similarly: stained first using immunofluorescence labels targeting WT1 (to generate training labels for podocyte detection), and then imaged via a whole-slide fluorescence scanner at 0.16 µm per pixel resolution and ×40magnification (Aperio Versa, Leica). These tissue sections were then post-stained using PAS, and bright-field imaged as described above. The human pathology WSIs used (Fig. 5) were obtained from 2–5-µm-thick biopsy sections, stained with PAS and bright-field imaged in a manner similar to that discussed above.

For digital MRI images of human prostate glands, 39 patients were recruited for an MRI scan befre a radical prostatectomy, using a 3T GE scanner (GE Healthcare) and an endorectal coil. The MRI included an axial T2-weighted image, collected with 3 mm slice thickness, 0.234 × 0.234 mm^2 voxel resolution, and a 4,750/123 ms TR/TE. The DICOM files were converted to NIFTI format using the mri_convert command from the Freesurfer library of tools (surfer.nmr.mgh.harvard.edu). Prostate masks were then manually annotated using AFNI by P.S.L. and verified by a board-certified radiologist for an unrelated study[47]. The prostate images and annotations were then converted into TIFF format using MATLAB (Mathworks Inc) for analysis by the SUNY Buffalo team.

In the H-AI-L pipeline, an annotator labels a limited number of WSIs using annotation tools in ImageScope[26], which provides the input for network training. The resulting trained network is then used to predict the annotations on new WSIs. These predictions are used as rough annotations, which are corrected by the annotator and sent back for incorporation into the training set; improving network performance and optimizing the amount of expert annotation time required. As this technique makes the adaptation of network parameters to new data easy, adapting a trained network to new data generated in different institutions is extremely feasible.

At the heart of H-AI-L is the conversion between mask and XML[48] formats, which are used by DeepLab v2[25] and ImageScope[26], respectively. Training any semantic segmentation architecture relies on pixel-wise image annotations that are input to the network for training and output after network predictions as mask images. In the case of DeepLab, the mask images take the form of indexed greyscale 8-bit PNG files, where each unique value pertains to an image class. On the other hand, annotations performed in ImageScope are saved in text format, as XML files[48], where each region is saved as a series of boundary points or vertices. Determining the vertices of a mask image is a common image processing task, known as image contour detection[49,50]. As opposed to edge detection, contour detection can have hierarchal classifications[50], lending itself ideally to conversion into the hierarchal XML format used by ImageScope.

To facilitate the transfer between ImageScope XML and greyscale mask images, we use the OpenCV-Python library (cv2)[49], specifically the function cv2.findContours to convert from masks to contours. Using this function, we are able to automatically convert DeepLab predictions to XML format, which can be viewed in ImageScope, and thus easily evaluate and correct network performance. Furthermore, we have written a library for converting an XML file into mask regions, using cv2.fillPoly. This library follows the OpenSlide-Python[51] conventions for reading WSI regions, returning a specified mask region from the WSI.

Using OpenSlide[51] and our XML to mask libraries allows for efficient chopping of WSIs into overlapping blocks for network training and prediction; similar sliding-window approaches are common in predicting semantic segmentations on large medical images[52,53]. To simplify the iterative training process, and complement the easy annotation pipeline proposed, we have created a callable function that handles operations automatically, prompting the user to initiate the next step. This function needs two flags [--option] and [--project], which are the parameters identifying the iterative step and the project to train, respectively. Initially created using [--option] 'new', a new project is trained iteratively by alternating the [--option] flag between 'train' and 'test'.

Multi-pass approach. Our algorithm uses our multi-pass approach by default. This approach is inspired by the way that pathologists scan WSIs at progressively higher resolutions. This process is accomplished by training two DeepLab segmentation networks using image regions and masks cropped from the training set. A high-resolution and a separate low-resolution network are respectively trained with full-resolution and down-sampled cropped regions. Prediction using this approach is performed serially; the low-resolution network identifies WSI regions to be passed to the high-resolution network for further refinement. This method is outlined in Supplementary Fig. 1.

Full-resolution analysis alone is achievable by setting the [--one_network] flag to 'True' during training and prediction. This analysis trains only the high-resolution network, which is exclusively used to segment WSIs during prediction. More information on the training and prediction is explained below.

Training. To streamline the training process, we created a pipeline where a user places new WSIs and XML annotations in a project folder structure, and then calls a function to train the project. This automatically initiates data chopping and augmentation, and then loads parameters from the most recently trained network (if available) before starting to train. For faster convergence, we utilize transfer learning, automatically pulling a pre-trained network file whenever a new project is created, which is used to initialize the network parameters before training. We have also included functionality to specify a pre-trained file from an existing project using the [--transfer] flag. For ease of use, the network hyper-parameters can be changed using command line flags, but are set automatically by default.

When [--option] 'train' is specified, WSIs and XML annotations are chopped into a training set containing 500 × 500 blocks with 50% overlap. This training set is then augmented via random flipping, hue and lightness shifts, and piecewise affine transformations, all accomplished using the imgaug Python library[54]. To keep the network unbiased, the total number of blocks containing each class is tabulated and used to augment less frequent classes with a higher probability[55]. Our multi-pass approach performs these steps for both high- and low-resolution patches separately to generate two training sets. The 500 × 500 low-resolution patches cover a greater receptive field, emphasizing information that occurs in the lower spatial image frequencies.

Once the training data have been assembled, the networks are trained for the specified number of epochs. The user is then prompted to upload new WSIs and run the [--option] 'predict' flag. This produces XML predictions that can be corrected using ImageScope before incorporation into the training set.

Multi-pass prediction. Due to the sparse nature of the structures we attempt to segment from renal WSIs, we limit the search space, using a low-resolution pass to determine hotspot regions before segmentation at full resolution. In this multi-pass approach, thresholding and morphological processing first determine which WSI blocks contain tissue, eliminating background regions. Second, down-sampled blocks (1/16th resolution, 500 × 500 pixels with 50% overlap) are extracted and tested, using the low-resolution segmentation network to roughly segment structures. The output predictions of the preprocessing steps are then stitched back into a hotspot map, which is 1/16th the WSI size. For multi-class cases, this stitching can be performed by finding the maximum class number between overlapping prediction maps, which is assigned to each pixel in the hotspot map. In this way, multi-class hierarchies are defined by assigning subclasses to higher mask indices. For example, conducting the stitching for the nuclear segmentation in Fig. 4 requires the definition of background, glomeruli, nuclei and podocyte classes to be 0, 1, 2 and 3, respectively, where nuclei and podocytes are compartments of glomeruli. The result in Fig. 5 was obtained using a similar procedure. This stitching operation is outlined in Supplementary Fig. 2 for two classes. The results in Figs. 2, 3 and 6 were obtained using a similar two-class stitching operation.

The hotspot map is then used to determine the locations for performing pixel-wise segmentation using the high-resolution DeepLab network (trained using full-resolution image patches). Hotspot indices are calculated, scaled back to full resolution (×16), and used to extract these regions at full resolution. The XML annotation file is then assembled from the high-resolution predictions on these regions.

Full-resolution prediction. When the [--one_network] flag is set to *'True'*, the initial extraction of overlapping blocks is performed at full resolution. Prediction

on these blocks uses the high-resolution DeepLab network, and the resulting hotspot map is stitched using the same method as above. Unlike above, this map (which is the same size as the WSI) is used to directly assemble the XML annotation file.

Post prediction processing. To limit possible false-positive predictions of small regions, we implemented a size threshold that tests the area of each predicted region, eliminating regions smaller than the set threshold using morphological operations. This threshold can be adjusted via the [--min_size] flag, and is easily estimated using the area displayed in the Annotations tab in ImageScope to determine the minimum regions size. By default, this threshold is set to 625 pixels, which was used for the analysis in this paper.

Validation. While the performance of the network is easily visualized after prediction on new WSIs, we have included functionality for explicitly evaluating performance metrics and prediction time on a holdout dataset. This is accomplished using the [--option] 'validate' flag. When called, it evaluates the network performance on holdout images for every annotation iteration by automatically pulling the latest models. To perform this performance comparison, ground-truth XML annotations of the holdout set are required to calculate the sensitivity, specificity, accuracy and precision performance metrics[38].

Estimating H-AI-L performance (Fig. 7). To quantify the time-savings of our H-AI-L method, we plot the normalized annotation time per region versus the number of regions annotated. Here we define the normalized annotation time per region A as $A=\frac{t}{t_0}$, where t is the annotation time per region (averaged per WSI) and t_0 is the average annotation time per region in iteration 0. A is bounded from [0,1], where 1 is the normalized time required to annotate one region fully. Although the annotation time is reduced as a piecewise function of the training iteration, in Fig. 7 we use a continuous exponential decay distribution to approximate $A(r)$:

$A(r)=e^{\frac{-r}{\tau}}$, where r is the number of regions annotated and τ is the exponential time constant, which we call the H-AI-L factor.

The normalized annotation time of our H-AI-L method (H) can therefore be estimated as

$$H=\int_0^R A(r)\mathrm{d}r=\tau[1-e^{\frac{-R}{\tau}}]$$

where R is the total number of regions annotated. Likewise, the normalized baseline annotation time (B) can be calculated as

$$B=\int_0^R 1\mathrm{d}r=R$$

Therefore, the time-savings performance (P) of our H-AI-L method can be estimated as a percentage:

$$P=\left(1-\frac{H}{B}\right)\times 100=\left(1+\frac{\tau}{R}[e^{\frac{-R}{\tau}}-1]\right)\times 100$$

The H-AI-L factor τ reflects the effectiveness of iterative network training, where lower values of τ represent training curves that decay faster. In the future, algorithms to select the optimal amount of annotation and identify data outliers to be annotated at each iteration will improve the performance of the H-AI-L method by reducing τ.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

We have made the data used for analysing the performance of H-AI-L method available at https://goo.gl/cFVxjn. The folder contains a detailed note describing the data. Namely, the folder contains pathology and radiology image data used for training and testing our H-AI-L method, ground-truth and predicted segmentations of the test image data, network corections and respective annotations of the training image data for different iterations, and the network models trained at different iterations. We have made our code openly available online at https://github.com/SarderLab/H-AI-L.

Received: 12 October 2018; Accepted: 7 January 2019;
Published online: 11 February 2019

References

1. Krizhevsky, A., Sutskever, I. & Hinton, G. E. ImageNet classification with deep convolutional neural networks. *Commun. ACM* **60**, 84–90 (2017).
2. LeCun, Y. & Bengio, Y. in *The Handbook of Brain Theory and Neural Networks* (ed. Michael, A. A.) 255–258 (MIT Press, Cambridge, 1998).
3. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436–444 (2015).
4. Pedraza, A. et al. Glomerulus classification with convolutional neural networks. In *Proc. Medical Image Understanding and Analysis: 21st Annual Conference, MIUA 2017* (eds Valdés Hernández, M. & González-Castro, V.) 839–849 (Springer, 2017).
5. Schmidhuber, J. Deep learning in neural networks: an overview. *Neural Netw.* **61**, 85–117 (2015).
6. Bottou, L. Large-scale machine learning with stochastic gradient descent. In *Proc. COMPSTAT'2010* (eds Lechevallier, Y. & Saporta, G.) 177–186 (Springer, 2010).
7. Szegedy, C. et al. Going deeper with convolutions. In *IEEE Conference on Computer Vision and Pattern Recognition* (IEEE, 2015).
8. Swingler, K. *Applying Neural Networks: A Practical Guide* (Morgan Kaufmann, Burlington, 1996).
9. Ronneberger, O., Fischer, P. & Brox, T. U-net: convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* (eds Navab, N., Hornegger, J., Wells, W. M. & Frangi, A. F.) (Springer, 2015).
10. Zhang, T. & Nakamura, M. Neural network-based hybrid human-in-the-loop control for meal assistance orthosis. *IEEE Trans. Neural Syst. Rehabil. Eng.* **14**, 64–75 (2006).
11. Krogh, A. & Vedelsby, J. in *Advances in Neural Information Processing Systems* (1995).
12. Cohn, D., Atlas, L. & Ladner, R. Improving generalization with active learning. *Mach. Learn.* **15**, 201–221 (1994).
13. Gosselin, P. H. & Cord, M. Active learning methods for interactive image retrieval. *IEEE Trans. Image Process.* **17**, 1200–1211 (2008).
14. Shi, L. & Wang, X.-c. Artificial neural networks: current applications in modern medicine. In *Computer and Communication Technologies in Agriculture Engineering, 2010 International Conference* (IEEE, 2010).
15. Madabhushi, A. & Lee, G. Image analysis and machine learning in digital pathology: challenges and opportunities. *Med. Image Anal.* **33**, 170–175 (2016).
16. Baxevanis, A. D. & Bateman, A. The importance of biological databases in biological discovery. *Curr. Protoc. Bioinformatics* **50**, 1.1.1-8 (2015).
17. Cheplygina, V. et al. in *Deep Learning and Data Labeling for Medical Applications* 209–218 (Springer, New York, 2016).
18. Szolovits, P., Patil, R. S. & Schwartz, W. B. Artificial intelligence in medical diagnosis. *Ann. Intern. Med.* **108**, 80–87 (1988).
19. Orthuber, W. et al. Design of a global medical database which is searchable by human diagnostic patterns. *Open Med. Inform. J.* **2**, 21 (2008).
20. Smeulders, A. W. et al. Content-based image retrieval at the end of the early years. *IEEE Trans. Pattern Anal. Mach. Intell.* **22**, 1349–1380 (2000).
21. Müller, H. et al. A review of content-based image retrieval systems in medical applications—clinical benefits and future directions. *Int. J. Med. Inform.* **73**, 1–23 (2004).
22. Gong, T. et al. *Automatic pathology annotation on medical images: a statistical machine translation framework*. In *Proc. 20th International Conference on Pattern Recognition* (IEEE, 2010).
23. Abe, N., Zadrozny, B. & Langford, J. *Outlier detection by active learning*. In *Proc. 12th ACM SIGKDD International Conference on Knowledge discovery and Data mining* (ACM, 2006).
24. Doyle, S. & Madabhushi, A. *Consensus of Ambiguity: Theory and Application of Active Learning for Biomedical Image Analysis* (Springer, Berlin, 2010).
25. Chen, L.-C. et al. Deeplab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. *IEEE Trans. Pattern Anal. Mach. Intell.* **40**, 834–848 (2018).
26. Aperio Imagescope (Leica Biosystems); https://www.leicabiosystems.com/digital-pathology/manage/aperio-imagescope/
27. Skodras, A., Christopoulos, C. & Ebrahimi, T. The JPEG 2000 still image compression standard. *IEEE Signal Process. Mag.* **18**, 36–58 (2001).
28. Sedeen Viewer (Pathcore); https://pathcore.com/sedeen/
29. Ginley, B., Tomaszewski, J. E. & Sarder, P. Automatic computational labeling of glomerular textural boundaries. In *Proc. SPIE 10140, Medical Imaging 2017: Digital Pathology* 101400G (2017).
30. Kato, T. et al. Segmental HOG: new descriptor for glomerulus detection in kidney microscopy image. *BMC Bioinformatics* **16**, 316 (2015).
31. Sarder, P., Ginley, B. & Tomaszewski, J. E. Automated renal histopathology: digital extraction and quantification of renal pathology. In *Proc. SPIE 9791, Medical Imaging 2016: Digital Pathology* 97910F (2016).
32. Simon, O., Yacoub, R., Jain, S., Tomaszewski, J. E. & Sarder, P. Multi-radial LBP features as a tool for rapid glomerular detection and assessment in whole slide histopathology images. *Sci. Rep.* **8**, 2032 (2018).
33. Tesch, G. H & Allen, T. J. Rodent models of streptozotocin-induced diabetic nephropathy. *Nephrology* **12**, 261–216 (2007).
34. Goyal, S. N. et al. Challenges and issues with streptozotocin-induced diabetes - a clinically relevant animal model to understand the diabetes pathogenesis and evaluate therapeutics. *Chem. Biol. Interact.* **244**, 49–63 (2016).
35. Kitada, M., Ogura, Y. & Koya, D. Rodent models of diabetic nephropathy: their utility and limitations. *Int. J. Nephrol. Renov. Dis.* **9**, 279–290 (2016).
36. Wu, K. K. & Huan, Y. Streptozotocin-induced diabetic models in mice and rats.*Curr. Protoc. Pharmacol.* **40**, 5.47 (2008).

37. Hripcsak, G. & Rothschild, A. S. Agreement, the F-measure, and reliability in information retrieval. *J. Am. Med. Inform. Assoc.* **12**, 296–298 (2005).
38. Sokolova, M., Japkowicz, N. & Szpakowicz, S. Beyond accuracy, F-score and ROC: a family of discriminant measures for performance evaluation. In *Australasian Joint Conference on Artificial Intelligence* (eds Sattar, A. & Kang, B.-H.) (Springer, 2006).
39. Japkowicz, N. & Stephen, S. The class imbalance problem: A systematic study. *Intell. Data Anal.* **6**, 429–449 (2002).
40. Bariety, J. et al. Parietal podocytes in normal human glomeruli. *J. Am. Soc. Nephrol.* **17**, 2770–2780 (2006).
41. Pavenstadt, H., Kriz, W. & Kretzler, M. Cell biology of the glomerular podocyte. *Physiol. Rev.* **83**, 253–307 (2003).
42. Solez, K. et al. Banff 07 classification of renal allograft pathology: updates and future directions. *Am. J. Transplant.* **8**, 753–760 (2008).
43. Mengel, M. Deconstructing interstitial fibrosis and tubular atrophy: a step toward precision medicine in renal transplantation. *Kidney Int.* **92**, 553–555 (2017).
44. Wang, X. et al. Glomerular pathology in dent disease and its association with kidney function. *Clin. J. Am. Soc. Nephrol.* **11**, 2168–2176 (2016).
45. McGarry, S. D. et al. Radio-pathomic maps of epithelium and lumen density predict the location of high-grade prostate cancer. *Int. J. Radiat. Oncol. Biol. Phys.* **101**, 1179–1187 (2018).
46. Janowczyk, A. et al. A resolution adaptive deep hierarchical (RADHicaL) learning scheme applied to nuclear segmentation of digital pathology images. *Comput. Methods Biomech. Biomed. Eng. Imaging Vis.* **6**, 270–276 (2016).
47. McGarry, S. D. et al. Radio-pathomic maps of epithelium and lumen density predict the location of high-grade prostate cancer. *Int. J. Radiat. Oncol. Biol. Phys.* **101**, 1179–1187 (2018).
48. Bray, T. et al. Extensible markup language (XML). *World Wide Web J.* **2**, 27–66 (1997).
49. Bradski, G. The OpenCV Library. *Dr. Dobb's* http://www.drdobbs.com/open-source/the-opencv-library/184404319 (2000).
50. Klette, R. et al. *Computer Vision* (Springer, New York, 1998)
51. Goode, A. et al. OpenSlide: a vendor-neutral software foundation for digital pathology. *J. Pathol. Inform.* **4**, 27 (2013).
52. Lu, C. & Mandal, M. Automated segmentation and analysis of the epidermis area in skin histopathological images. In *2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society* (IEEE, 2012).
53. Govind, D. et al. Automated erythrocyte detection and classification from whole slide images. *J. Med. Imaging* **5**, 027501 (2018).
54. Jung, A. imgaug (2017); http://imgaug.readthedocs.io/en/latest/
55. Zhou, Z.-H. & Liu, X.-Y. Training cost-sensitive neural networks with methods addressing the class imbalance problem. *IEEE Trans. Knowl. Data Eng.* **18**, 63–77 (2006).

Acknowledgements

The project was supported by the faculty start-up funds from the Jacobs School of Medicine and Biomedical Sciences, University at Buffalo, the University at Buffalo IMPACT award, NIDDK Diabetic Complications Consortium grant DK076169 and NIDDK grant R01 DK114485. The prostate imaging data were collected with funds from the State of Wisconsin Tax Check-off Program for Prostate Cancer research. Percent efforts for P.S.L. and S.D.M. were provided by R01 CA218144, and the National Center for Advancing Translational Sciences NIH UL1TR001436 and TL1TR001437. We thank NVIDIA Corporation for the donation of the Titan X Pascal GPU used for this research.

Author contributions

B.L. conceived the H-AI-L method, analysed the data and wrote the manuscript. The code was written by B.L. and B.G. D.G. contributed in generating results for Fig. 4. S.D.M. and P.S.L. provided the radiology data and annotations for the prostate MRI analysis, and edited the manuscript. R.Y. implemented the mouse model. S.J. provided human renal biopsy data. J.E.T. evaluated renal pathology segmentation as a domain expert. K.-Y.J. provided the IFTA annotation for Fig. 5. P.S. is responsible for the overall coordination of the project, mentoring and formalizing the image analysis concept and oversaw manuscript preparation.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s42256-019-0018-3.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to P.S.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature Limited 2019

nature research

Corresponding author(s): Pinaki Sarder

Last updated by author(s): Dec 24, 2018

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☒	☐	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☒	☐	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☒	☐	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	☐	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Leica digital pathology scanners were used for the collection of the digital whole slide images (WSI) of the histology specimens used in this study. Digital MRI image collection was standard and is described in the Methods section of the manuscript.
Data analysis	Aperio Imagescope version 12.3.3 (Leica) was used for digital image annotation and viewing. The Openslide python library was used to view digital images in python. The H-AI-L code is written in python 3. The semantic segmentation network, DeepLab v2 was used for this work.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

We have made the data used for the performance analysis available here: https://goo.gl/cFVxjn. We have made our code openly available online: https://github.com/SarderLab/H-AI-L

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	NA
Data exclusions	No data were excluded
Replication	Training of neural networks with large datasets takes time, which limited the possibilities for replication. However, during development, we were able to obtain similar performance to that presented in this manuscript using random testing experiments.
Randomization	The samples were randomly assigned to training and holdout testing sets for all analysis done in this manuscript. During training the inclusion of new training data was done by selecting WSIs where the network performed poorly (done to optimize the learning speed).
Blinding	Experts performing annotation for this study were not given information on the disease or condition of the data they were given. This was done to ensure unbiased annotation performance.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☐	☒ Animals and other organisms
☐	☒ Human research participants
☒	☐ Clinical data

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Animals and other organisms

Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research

Laboratory animals	Mouse, C57BL/6J background, Male, 7-32 weeks of age.
Wild animals	NA
Field-collected samples	NA
Ethics oversight	Institutional Animal Care and Use Committee at University at Buffalo

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Human research participants

Policy information about studies involving human research participants

Population characteristics	18-70 years old. Special populations (vulnerable) such as minors, pregnant women, neonates, prisoners, children, and cognitively impaired patients were not included. There are no restrictions in regards to gender or race. The renal tissue samples were obtained at Kidney Translational Research Center of Washington University School of Medicine under their institutionally approved protocol. Tissue samples were provided to the corresponding author and the research team after de-identification under an institutionally approved protocol at University at Buffalo. The human prostate MRI digital images were provided by co-author Peter S. Laviolette to the research team after de-identification under an institutionally approved protocol at Medical

	College of Wisconsin.
Recruitment	NA. Digital images were used in a de-identified fashion for developing the computational image annotation pipeline. For achieving robustness, several images from several cases were needed. However, any specific recruitment strategy does not impact the computational result produced.
Ethics oversight	Washington University School of Medicine, University at Buffalo, Medical College of Wisconsin.

Note that full information on the approval of the study protocol must also be provided in the manuscript.