Learning fast and fine-grained detection of amyloid neuropathologies from coarse-grained expert labels

Daniel R. Wong[1,2,3,4,5], Shino D. Magaki[6], Harry V. Vinters[6,7], William H. Yong[8], Edwin S. Monuki[8], Christopher K. Williams[6], Alessandra C. Martini[8], Charles DeCarli[9], Chris Khacherian[8], John P. Graff[10], Brittany N. Dugger[10*], Michael J. Keiser[1,2,3,4,5*]

1. Institute for Neurodegenerative Diseases, University of California, San Francisco, San Francisco, CA, 94158, USA
2. Bakar Computational Health Sciences Institute, University of California, San Francisco, CA, 94158, USA
3. Department of Pharmaceutical Chemistry, University of California, San Francisco, San Francisco, CA, 94158, USA
4. Department of Bioengineering and Therapeutic Sciences, University of California, San Francisco, San Francisco, CA, 94158, USA
5. Kavli Institute for Fundamental Neuroscience, University of California, San Francisco, San Francisco, CA, 94158, USA
6. Department of Pathology and Laboratory Medicine, University of California, Los Angeles, Los Angeles, CA, 90095, USA
7. Department of Neurology, David Geffen School of Medicine at University of California, Los Angeles, Los Angeles, CA, 90095, USA
8. Department of Pathology & Laboratory Medicine, University of California, Irvine, CA 92697, USA
9. Department of Neurology, School of Medicine, University of California-Davis, Davis, CA 95817, USA
10. Department of Pathology and Laboratory Medicine, School of Medicine, University of California, Davis, Sacramento, CA 95817, USA

*Correspondence: bndugger@ucdavis.edu, keiser@keiserlab.org

Abstract

Precise, scalable, and quantitative evaluation of whole slide images is crucial in neuropathology. We release a deep learning model for rapid object detection and precise information on the identification, locality, and counts of cored plaques and cerebral amyloid angiopathies (CAAs). We trained this object detector using a repurposed image-tile dataset without any human-drawn bounding boxes. We evaluated the detector on a new manually-annotated dataset of whole slide images (WSIs) from three institutions, four staining procedures, and four human experts. The detector matched the cohort of neuropathology experts, achieving 0.64 (model) vs. 0.64 (cohort) average precision (AP) for cored plaques and 0.75 vs. 0.51 AP for CAAs at a 0.5 IOU threshold. It provided count and locality predictions that correlated with gold-standard CERAD-like WSI scoring ($p=0.07\pm 0.10$). The openly-available model can quickly score WSIs in minutes without a GPU on a standard workstation.

Introduction

Deep phenotyping of Alzheimer's disease requires accurate evaluation of whole slide image (WSI) data[1]. For amyloid-β (Aβ) pathologies, such as plaques and cerebral amyloid angiopathy (CAA), quantifying Aβ burden phenotypes in the brain can aid in understanding disease mechanisms and progression[2–4]. However, in neuropathological practice, quantifying Aβ burden has primarily been semi-quantitative[5] with variable interpretation[6,7]. Interpreting WSIs is a time-consuming task[8] with pathologists regularly spending many hours a day assessing slides[9].

Deep learning has helped to address these challenges, providing quantitative and automated solutions to identifying and quantifying Aβ burden[7,10,11]. Deep learning can augment neuropathologist expertise[10] and combine multiple expert annotations into a robust and automated labeler[7]. For more localized tasks like object detection[12] and semantic segmentation[13], deep learning has also provided accurate and automated means of quantifying Aβ and tau neuropathologies[14,15,16]. However, such studies require significant human expert labor to create high-quality training datasets in the form of manually drawn bounding boxes or segmentations and categorical labels. Furthermore, the models typically require specialized dedicated and expensive hardware like graphic processing units (GPUs)[17], without which the prediction task of quantifying pathologies can take hours for even a single WSI. Furthermore, as with many deep learning studies, generalizability to data from different institutions is difficult to guarantee[11,18,19].

Here, we present a fast You Only Look Once version three (YOLOv3) based model[20] that rivals human-expert level detection of cored plaque and CAA pathologies. Moreover, we created this model from a dataset not intended for object detection, requiring much less human labor than a traditional object detection dataset. We evaluated the model on WSIs outside of its training corpus, which were diverse in both stain and institutional source. The model, released at https://github.com/keiserlab/amyloid-yolo-paper, can quickly score WSIs without a GPU, paving the way for more accessible and equitable deep learning applications in the research and clinical space. Furthermore, we showed that without a GPU, the model can still score WSIs in a matter of minutes, with speed improvements of at least eight times over various state-of-the-art deep learning approaches for quantifying neuropathologies[10,16]. To determine the model's potential for

adoption of widespread scoring use, we evaluated it on WSIs with known CERAD-like scores[5] and found strong correspondence. The model enables scalable, reproducible, and precise detection for rapid clinical research applications.

Results

We built an object detector from a noisy and sparse dataset

We repurposed a dataset from a previous study[7] not meant for object detection training. The previous study had collected human annotations post hoc on a 256 x 256 pixel tile basis for tiles centered on approximate bounding boxes of cored and CAA pathologies derived from traditional and automated computer vision techniques (Methods). Consolidating this dataset into 659 larger field-of-view (1536 x 1536 pixel) images devoid of human-drawn boxes, we reformatted the data to a form more suitable for object detection. This dataset had many limitations: 1) a single pathology often incorrectly spanned many approximate boxes, particularly for CAAs (Supplemental Figure 1); 2) the 1536 x 1536 fields lacked comprehensive annotations, resulting in a sparse label set prone to false negatives; 3) traditional watershed techniques defined each box, rather than human intelligence; 4) the dataset size was relatively small (659 images from 29 WSIs); and 5) due to limitations 1, 2, and 3 there was no reliable quantitative benchmark to assess model performance. For this study, we did not collect further human annotations for training, instead adapting the existing dataset to a more suitable form. Limitations 2-5 would have required more human annotation work. To solve limitation 1, we performed an iterative merging procedure such that overlapping label boxes of the same class were joined (Supplemental Figure 1; Methods).

Once we merged the label data for use in the current study, we trained a YOLOv3 network[20] to identify cored plaque and CAA pathologies (Methods). We denote this initial model as model-1. Average precision (AP) for each class at varying intersection-over-union (IOU) thresholds typically exceeded 0.6 (Supplemental Figure 2); however, the model had some flaws. Visually, model-1 incorrectly labeled single instances with many overlapping boxes, especially for CAAs (Supplemental Figure 2). Hence, we trained a new model incorporating two enhancements. For the first, we joined overlapped output CAA predictions from model-1 to

consolidate the fragments (Methods). Secondly, we used our previously released consensus-of-two convolutional neural network (CNN) model[7] to filter out low-quality CAA detections. Finally, we merged output boxes of the same class, arriving at our final model, denoted model-2 (Methods). Figure 1 shows model-2's average precision over varying IOU thresholds (Supplemental Figure 3) and example image predictions. Although model-2's average precision over varying IOU thresholds (Figure 1) does not greatly differ from model-1's, model-2 identified pathologies better by visual evaluation. This is sensible, as improved bounding box quality not only improved training but also increased the stringency of the validation benchmark. Consequently, we used model-2 for the remainder of the study.

Fine-grained human expert annotations of pathologies were variable

To assess the model's prospective capabilities, we needed a higher quality test dataset (one free from limitations 1-3) to derive reliable quantitative metrics. For this, four experts (anonymized as NP1-NP4) independently annotated an entirely new dataset from a new decedent cohort, drawing boxes around and classifying pathologies (Methods). These test data differed markedly from our training and validation data, which only had sparse and incomplete computer-generated boxes with expert labels. This new dataset consisted of 200 1536 x 1536 pixel images spanning four different immunohistochemical stains for amyloid beta deposits. We found that the four neuropathology annotators did not always agree on this fine-grained task, with average agreement accuracy for cored = 0.43 ± 0.05 and CAA = 0.33 ± 0.11 at an IOU threshold = 0.50 (Figure 2). Given this variability, we additionally created a "consensus annotation" benchmark set wherein each "positive" object and its box were independently supported by at least two out of the four annotators (Figure 2B; Methods).

The model achieved human-expert-level precision

When we assessed the model against both individual-expert annotations and the consensus annotation datasets, we found it achieved expert-level precision at identifying both cored plaques and CAA pathologies on these new datasets despite never having been trained on manual bounding boxes (Figure 3). To cross-compare expert consistency across the annotations and thereby determine the achievable performance range from human variability, we treated each expert's annotations as though they were model predictions and compared them against each

other. In this procedure, each annotator's labels sequentially became a ground truth benchmark, against which we compared every other expert's annotations; we subsequently calculated the average precision (shown as the blue-dotted line in Figure 3A). For most IOU thresholds (less than or equal to 0.70) and most benchmarks, the model operated within the range of human-expert level performance. For CAAs, the model's AP exceeded the average AP between human experts for four out of five benchmarks at IOU thresholds less than or equal to 0.60. For the strictest IOU thresholds ≥ 0.80, which require a more exact match between the predicted and label bounding box coordinates, the model fell short of human-expert performance (which itself was relatively low) for four out of five benchmarks. Of all the human experts, the model's predictions most closely matched the annotations of NP1, who spent significantly more time annotating than any of the other annotators (Figure 3A). NP1 spent nearly three times as long as NP2 and about twice as long as NP3 and NP4.

Model predictions correlated with clinical CERAD-like scores

We sought to test whether the model could quantify select amyloid-beta deposits, CAAs, and cored plaques on WSIs. Hence, we asked if an automated score calculated for entire WSIs based on the model's detection of amyloid pathologies would reflect human-expert-based CERAD-like category scoring[5]. We used a testing holdout set of 63 WSIs labeled with this semi-quantitative gold standard for pathology from a previous study[10].

We used our model to exhaustively detect and count pathologies within each of the 63 WSIs. We found that the counts derived from the model predictions significantly correlated with CERAD-like severities (Figure 4). We performed a two-sided student's t-test to determine if the model-derived count distributions differed significantly between the different CERAD-like categories. We found significant differences for all category pairs at an alpha of 0.05, except for the pair "none" and "sparse" (student's t-test p-value = 0.30, power = 0.62). Power for all other comparisons exceeded 0.99.

The model is much faster than existing approaches

Next, we evaluated the model's practical usability as measured by its speed. Hence, we evaluated the 63 WSIs from our CERAD-like dataset using one NVIDIA Titan Xp GPU and

computed the model's average speed per WSI. The model averaged one minute and forty seconds to score a single WSI. Consumer-grade GPUs are not universally available in pathology practices, so we tested model speed without GPU; the model averaged five minutes and thirty seconds per WSI on an Intel Xeon CPU (Supplemental Table 1).

We compared this YOLOv3 model's evaluation time with two different deep-learning approaches for quantifying neuropathology burden (Table 1). First, we compared with our previously published approach of using a CNN sliding window to count plaque burden, which took three hours and four minutes per WSI on an NVIDIA GTX 1080 GPU[10]. Even without a GPU, this YOLOv3 model was 33 times faster than the older GPU-*enabled* sliding window approach. On average, when we used a GPU, the YOLOv3 model was 110 times faster than our previous approach[21]. Second, we compared the YOLOv3 model's speed to a semantic segmentation method for tauopathies[16]. This different state-of-the-art approach to quantifying neuropathologies reported 45 minutes per WSI using a GPU; this YOLOv3 model was 8x-27x faster, depending on GPU usage. Exact runtimes will vary by hardware.

Discussion

We present a rapid object-detection model for identifying amyloid-beta cored plaques and cerebral amyloid angiopathy across a range of immunohistochemically stained slides. Three points of the study merit particular emphasis: 1) we developed a detector model from a dataset that did not require the same time and labor usually needed for building accurate object detectors; 2) the model matched human-expert performance; and 3) the model showed promise for usability without special GPU hardware for evaluation. Regarding the first point, we overcame one of the main problems with training object detectors through an iterative process: manual labeling and localization of high-quality bounding box data. The model still relied on accurate categorical labels in its training, but this is much less work than drawing a box and providing a categorical label, which is important for scalability. We hope our proof-of-concept encourages other deep learning studies to explore the prospect of bootstrapping from a more economical standpoint via more pragmatic proxy data, perhaps even by building more directly off of preliminary training data from conventional computer vision tools. We were encouraged to

see that the approach could identify the two pathologies with high precision, starting from only 659 initially noisy and sparsely-annotated high-resolution training images.

The final model achieved human-expert-level performance on prospective validation despite using a lower-quality training dataset devoid of human-derived bounding boxes. Although the study's scope of four expert annotators and 200 prospectively annotated images does not ensure generalizability to all experts, pathologies, stains, areas, cases, and institutions, it was encouraging to see that the model most closely aligned with the annotator who spent the most time annotating (Figure 3). With expert-level precision on held-out data, models such as these may readily be used as a secondary or preliminary labeler by neuropathologists, particularly to flag unusual cases. The current model's effectiveness will depend strongly on whether an expert needs strict down-to-the-pixel bounding-box overlap between the prediction and the actual pathology because the model falls slightly below human-expert level performance at the strictest IOU thresholds. For annotation tasks demanding high locality, a pixel-by-pixel level of precision via semantic segmentation may be a more apt technical approach.

Accordingly, several caveats apply to the study. First, although our prospective validation dataset was composed of cases across three different institutions, variable in stain, and annotated by four experts, it consisted of only two hundred 1536 x 1536 pixel images derived from 56 WSIs. Therefore much of the model's generalizability has yet to be explored. Likewise, performance on stains outside of the four used has yet to be determined, although we did not see much performance variation by stain except for 6E10 (Supplemental Figure 4). Furthermore, CAA pathology is diverse[22], but this model does not differentiate between subtypes; we would be interested to explore CAA-subtype identification in a future study. Finally, we relied on author-reported runtime analyses for various comparative speed benchmarks against alternative computational methods when their code was unavailable. These assessments necessarily spanned 1-2 generations of CPU and GPU hardware. However, given that a GPU is approximately two orders of magnitude faster at deep learning tasks than the contemporaneous CPU, we found the CPU-only speedup of the YOLOv3 model against GPU-enabled alternatives compelling.

The model's predicted amyloid-deposit counts correlated significantly with CERAD-like category scoring at the WSI level (Figure 4) without any training specifically for this purpose. A score of predicted-object counts struggled to significantly differentiate the "sparse" versus

"none" CERAD-like categories, perhaps due to the low sample sizes (n=6 and n=11) and resulting in a low power of 0.62. Nonetheless, we hope the model can quickly assess WSIs and provide a proxy for CERAD-like scoring, especially in detecting cases with higher plaque burden.

We freely release the trained model, annotated dataset, and study source code for easy access and use. We provide an example environment (using conda) to standardize model deployment. Although we found the consensus-annotation benchmark seemed a more stable label dataset than any given individual expert's annotations *on average*, consistent with the common notion of the "wisdom of the crowd," we did not tune the model for *specific* expertise nor neuropathology focus areas. Consequently, interested researchers may wish to fine-tune or entirely retrain this model on more precisely formulated annotations to fit their needs. We hope that the model and dataset's open-source release, enabling quick evaluation of WSIs even without a GPU, will facilitate the shareable and scalable application of deep learning in neuropathology.

Methods

Training and Validation Dataset Preparation

The training of model versions one and two builds on the methods and dataset first presented in Wong *et al.*[7] We provide a brief description of the methods used to build this dataset. We collected 29 WSIs of the temporal cortex from 3 different sites: 11 from the Alzheimer's Disease Center at the University of California, Davis (UC Davis); 11 from the University of Pittsburgh; and seven from UT Southwestern (Supplemental Figure 5). Slides were derived from formalin-fixed paraffin-embedded sections and stained with an antibody directed against Aβ. UC Davis used the Aβ 4G8 antibody, the University of Pittsburgh used a NAB228 antibody, and UT Southwestern used a 6E10 antibody. We imaged all WSIs on an Aperio AT2 at either 20X or 40X magnification at the different institutions. We resized all 40X images to 20X. For patient demographic data, please refer to Wong *et al.*[7]

We color-normalized the WSIs[23]. Each WSI was uniformly tiled to 1536 x 1536 pixel non-overlapping images. After tiling, we applied a hue saturation value (HSV) color filter and smoothing technique to detect candidate plaques using the python library openCV. We used different HSV ranges for the different stain types as follows: 4G8 HSV = (0, 40), (10, 255), (0, 220); NAB228 HSV = (0, 100), (1, 255), (0, 250); and 6E10 HSV = (0, 40), (10, 255), (0, 220).

Within each tiled 1536 x 1536 pixel image, each candidate pathology was bounding boxed via the watershed algorithm and then annotated by four neuropathology experts. Each expert performed a multi-class labeling task, selecting any or none of the classes: cored plaques, Diffuse, and CAA. After annotation, we applied a consensus-of-two strategy to obtain our final label set, such that a candidate plaque *p* was recorded as positive if any two experts marked *p* as positive for class *c*, else we recorded *p* as negative. We discarded the diffuse pathologies from our dataset and focused on the classes cored plaques and CAA. If any bounding box of class *c* overlapped with any other bounding box of class *c*, we merged the boxes into a single bounding box of class *c, which* was the minimal superset of the two bounding boxes. The result was 659 1536 x 1536 pixel images that contained either a cored or CAA pathology.

Training Model Version One

We split the 29-WSI dataset of 659 images into 70% training and 30% validation. We trained an initial YOLOv3 network from the image dataset for 200 epochs using a pre-trained Darknet, a batch size of eight, and a learning rate of 0.001. The file config/yolov3-custom.cfg at https://github.com/keiserlab/amyloid-yolo-paper contains full training hyperparameters. We selected the model weights from the epoch giving us the highest mean average precision over the validation set.

Training Model Version Two

We ran model version one over the training set. We merged model output prediction boxes of the same class using the same data preprocessing procedure as the original bounding box merging. We combined the resulting merged predictions with the existing training labels to create a new training dataset. We then used the model published in Wong et al.[7] to remove bounding boxes with predicted confidence < 0.5 for the relevant deposit to filter unlikely (false-

positive) model-1 predictions. We trained a new model from this new training dataset, called model version two. The training parameters were the same as those used for training model version one.

Selecting Images for Prospective Validation

To assess our model on data it had never seen, we collected a new dataset of 56 WSIs that differed from those used for training and validation. These new WSIs came from three different institutions: UC Davis, UC Los Angeles (UCLA), and UC Irvine (UCI). UC Davis used a 4G8 stain, UCLA used both an ABeta40 and ABeta42 stain, and UCI used a 6E10 stain (Supplemental File 1). We imaged all WSIs on an Aperio AT2 (UC Davis), Aperio CS2 (UCLA), and an Aperio Versa 200 (UCI) scanner at 20X or 40X magnification at the different institutions. For UCLA and UC Davis slides, each pixel corresponds to 0.5 microns. For UCI slides, each pixel corresponds to 0.274 microns. For each of the four stains, we selected the top 12 WSIs with the highest count of human-annotated CAAs, resulting in 48 different WSIs. For each of these 48 slides, we selected three 1536 x 1536 pixel fields to be used for prospective validation as follows: 1) field with the largest count of CAA positive model predictions (from model version two); 2) field with the largest count of CAA positive human annotations; and 3) top two fields with the largest count of cored positive model predictions (from model version two). Additionally, for each of the four stains, we randomly selected two WSIs that were different from the original 48. We randomly picked two fields for each of these eight WSIs. This resulted in a total prospective validation set size of 200 images, each with 1536 x 1536 pixels.

Annotating Prospective Validation Images

Four neuropathology experts independently annotated each of the 200 images used for prospective validation. These experts were different from the original five who helped to create our training and validation dataset in the prior study[7], except for one expert (B.N.D.) who helped with labeling the training and validation set and the prospective validation images. We provided standardized annotation instructions to all experts (Supplemental File 1). We used the web platform called "SuperAnnotate"[24] for obtaining bounding box labels. Each annotator had one month to complete the annotations.

Assessing Interrater Agreement

We evaluated the interrater agreement accuracy between two annotators (denoted as "A1" and "A2" in this section) for prospective validation as follows. For the superset of all pathologies of class P (either cored or CAA) identified by A1 or A2 (with cardinality "total"), we determined if both annotators gave congruous labels for each plaque. Two label boxes form a congruous pair if they share the same class label and intersect with IOU threshold of at least 0.50. We allowed each label pathology to be a part of at most one congruous pair (i.e., if multiple of A1's labels overlapped with a single label from A2, only one of A1's labels was part of the pair). We defined "overlaps" as the number of congruous pairs between A1 and A2. The final interrater accuracy between A1 and A2 derives from "overlaps" divided by "total" (Figure 2).

Assessing Model Performance on the Prospective Validation Images

We used model version two to assess performance on the prospective validation images. For each of the 200 images, we derived final predictions by merging any predicted bounding boxes that overlapped with any others of the same class. For each CAA bounding box prediction, we center-cropped the box to derive a 256 x 256 pixel image and fed this image into a previously published CNN model (the consensus-of-two model first presented in Wong et al.[7]). If the consensus-of-two model gave a negative CAA prediction, then this prediction was removed.

For all model evaluations, if there were multiple detections for a single label, we counted the highest confidence detection as a true positive and the rest as false positives (keeping consistent with the precedent set forth by the PASCAL VOC challenge[25]).

To determine the ceiling performance that we could expect from our model, we assessed how well each expert annotator matched the other experts (blue shaded region in Figure 3A). For each expert annotator (denoted as "A" for this section), we first fixed A as the ground truth, compared the other annotators' labels (not including A) to A's ground truth, and derived the precision. We did not include the consensus "annotator" in this comparison. We averaged the resulting 12 different comparisons and precision scores for each IOU threshold and plotted the standard deviation around the average (Figure 3A).

Correlating Model-derived Plaque Counts with CERAD-like Scoring

For each of the 63 WSIs with CERAD-like scores available in Tang *et al.*[10], we tiled the WSI into non-overlapping 1536 x 1536 pixel tiles. We ran model version two over all tiles to identify predicted pathologies. We merged any predicted bounding boxes that overlapped with any others of the same class and counted the number of predicted cored bounding boxes to compare with the clinical CERAD-like score.

We performed a two-sided student's t-test between each CERAD-like category's distribution of model-derived plaque counts. The null hypothesis was that the two distributions were no different, and the alternative hypothesis was that the two distributions were indeed different. Each point of any distribution was a single model-derived plaque count from one WSI. We used an alpha threshold of 0.05 to assign significance.

Figure Legends

Figure 1: Model version two performance and example image predictions. Top: Average precisions (AP) over the validation set at various IOU thresholds. The AP at IOU=0.90 is undefined for CAA. Bottom: 16 example images from the validation set. Cored prediction: red, cored label: black "*"; CAA prediction: blue, CAA label: black "@". Note that these training label data are sparse and do not contain every pathology (Methods).

Figure 2: Fine-grained human bounding-box style annotations vary slightly. (a) Interrater agreement accuracy among annotators, with a minimal IOU threshold of 0.50 used for counting two objects of the same class as an overlap (Methods). (b) Left column: example overlaid annotations from each of the four annotators (each a different color); Right column: corresponding consensus annotation.

Figure 3: Model achieved human-expert level performance at identifying cored and CAA pathologies. (a) Average model precision scores for identifying cored pathologies (left) and CAA pathologies (middle). Y-axis: average precision, x-axis: IOU threshold that determines the minimal IOU required for a prediction to overlap with a label to be a true positive. Higher IOU thresholds are more stringent. The figure legend indicates which of the annotators is the ground truth benchmark for assessing the model. The black line indicates model AP against the consensus annotator benchmark (Figure 2B, right column). The blue dotted line is the average precision of comparing expert annotators to each other (Methods). The blue-shaded region is one standard deviation above and below the average-expert precision. Right: total hours each annotator spent annotating (x-axis) versus AP at IOU=0.50 of the model on the annotator's benchmark (y-axis). "*" indicates cored performance, "@" indicates CAA performance. (b) Model predictions overlaid against consensus annotation. Cored prediction: red, cored label: black "*"; CAA prediction: blue, CAA label: black "@". The consensus annotation defines the labels.

Figure 4: Model correlated with clinical CERAD-like scoring. Left: box plots for each CERAD-like category. Y-axis is the model-derived count of cored plaques, and the x-axis is the CERAD-like category. Scatter plot overlaid as blue dots (each dot corresponds to a unique WSI). Hollow black circles indicate outliers outside the third quartile plus 1.5x interquartile range.

*$p<0.05$, n.s. is not statistically significant. Right: p-values from a two-sided student's t-test comparing model-derived count distributions between each CERAD-like category.

Figures

Figure 1

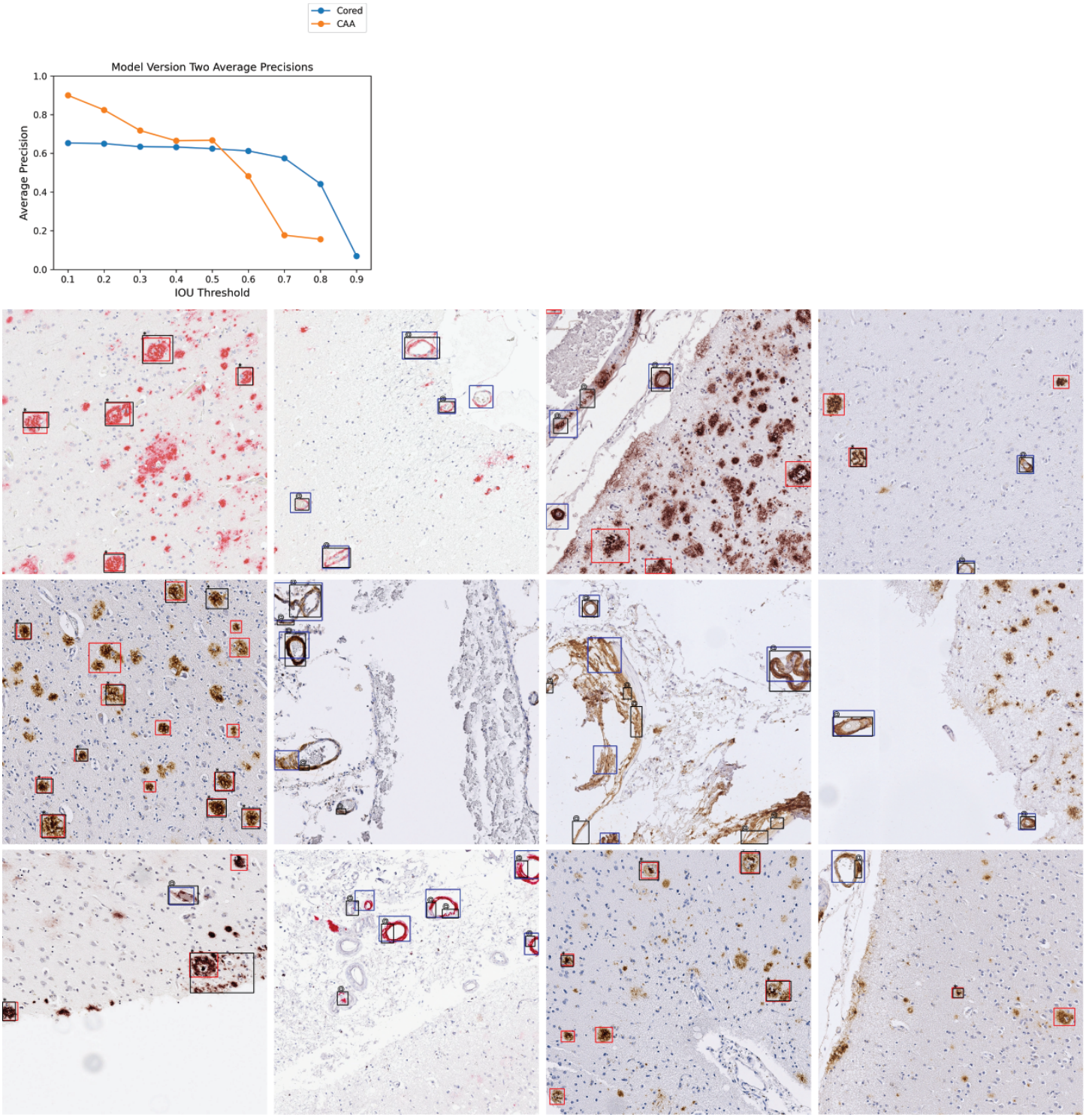
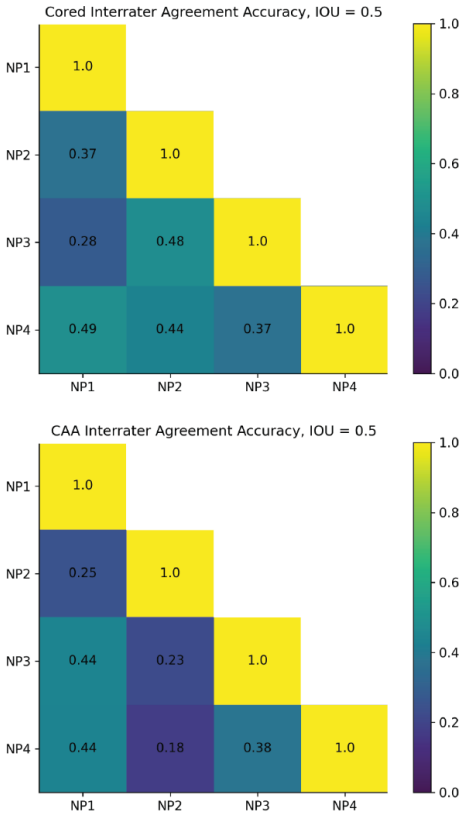


Figure 2

A



B

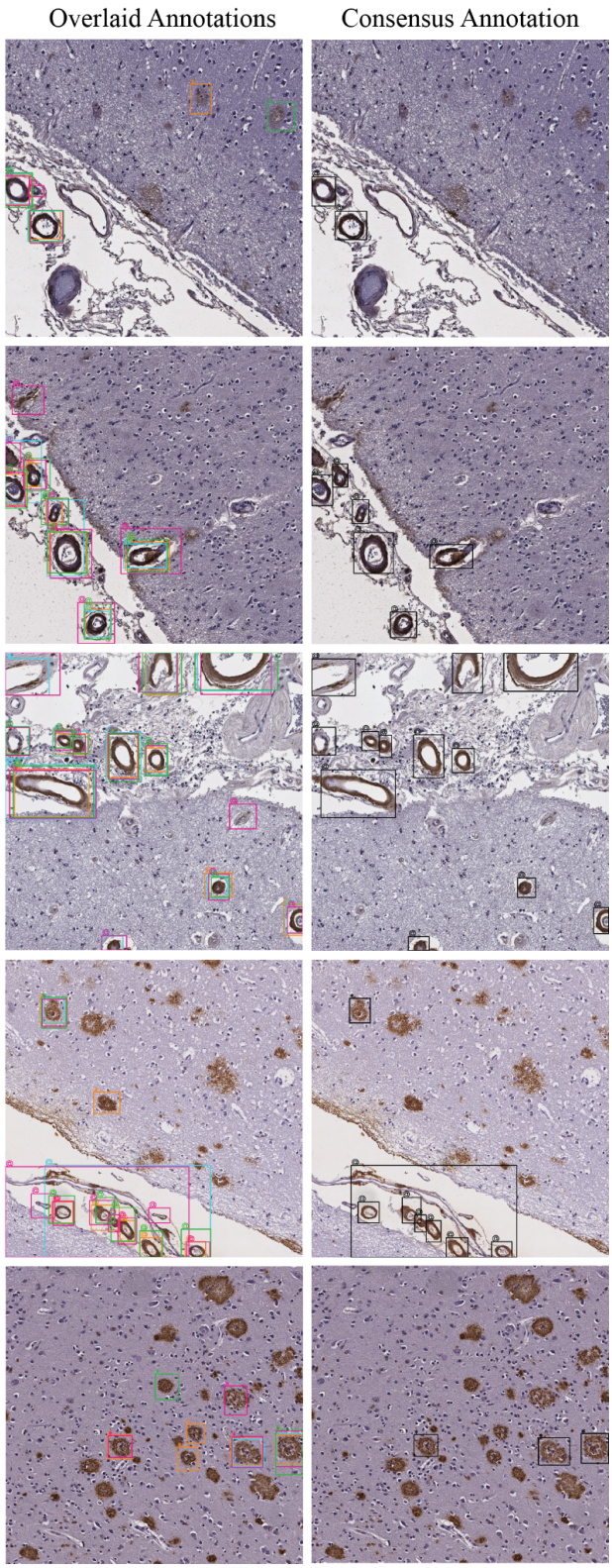
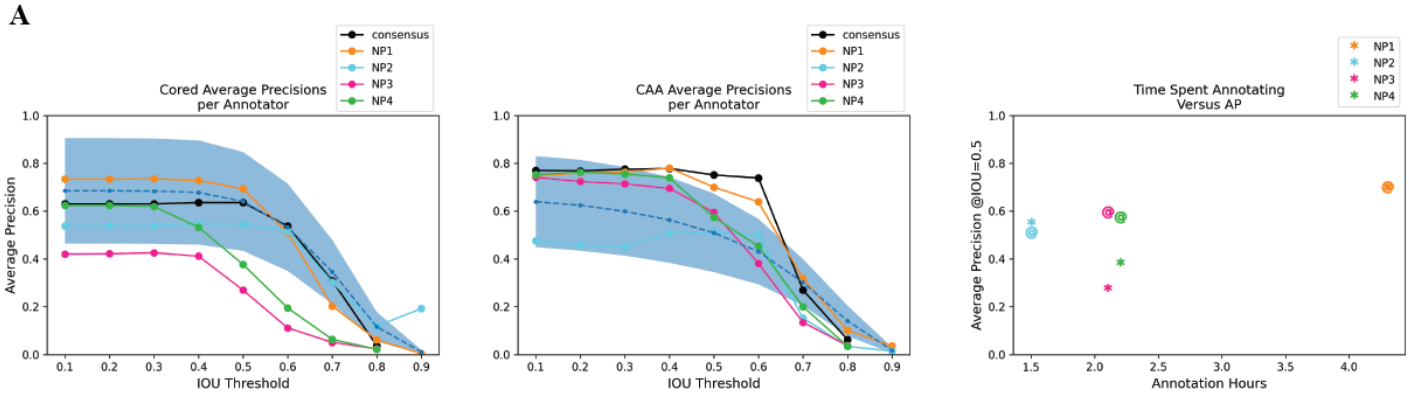


Figure 3

A



B

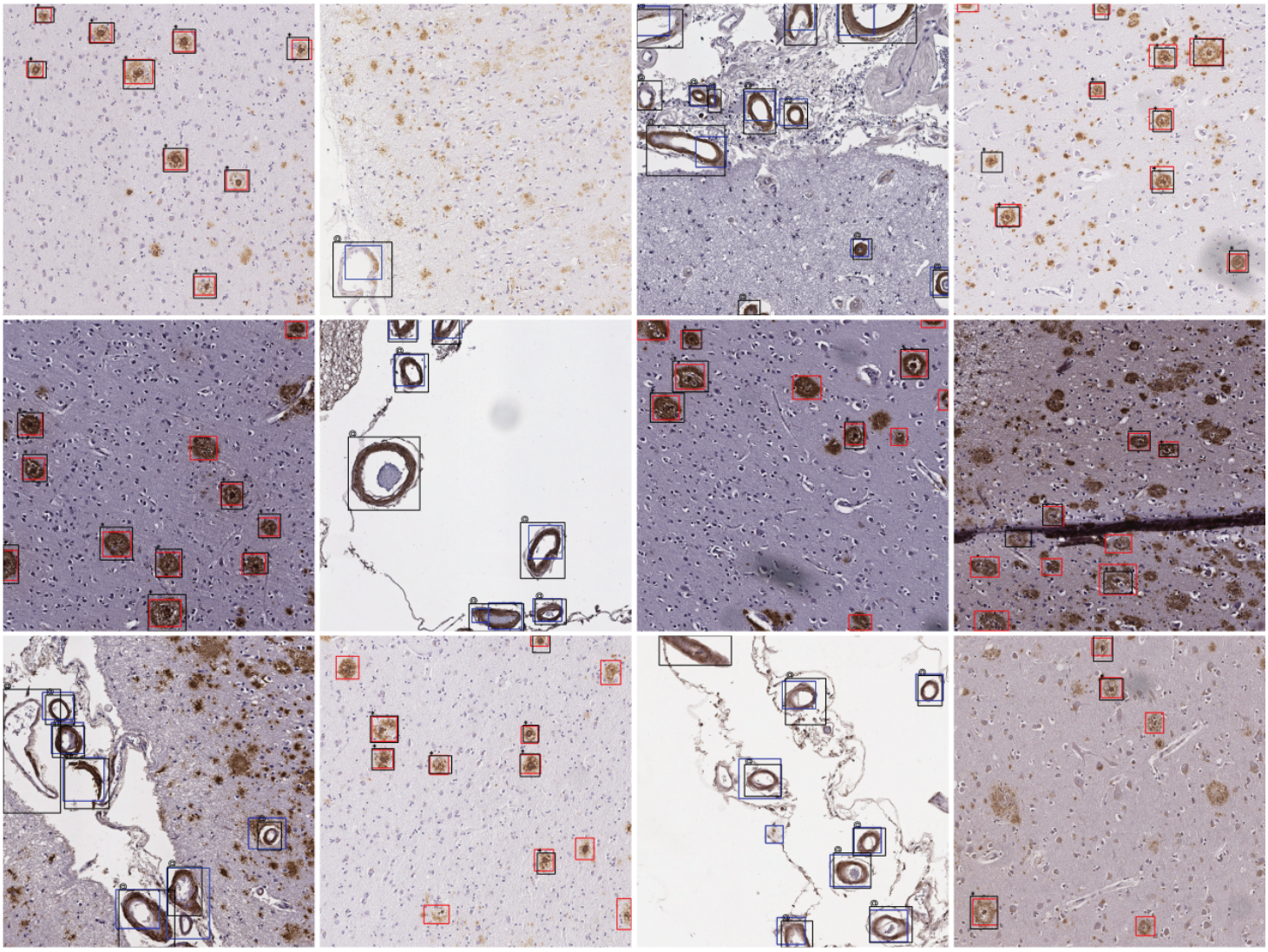
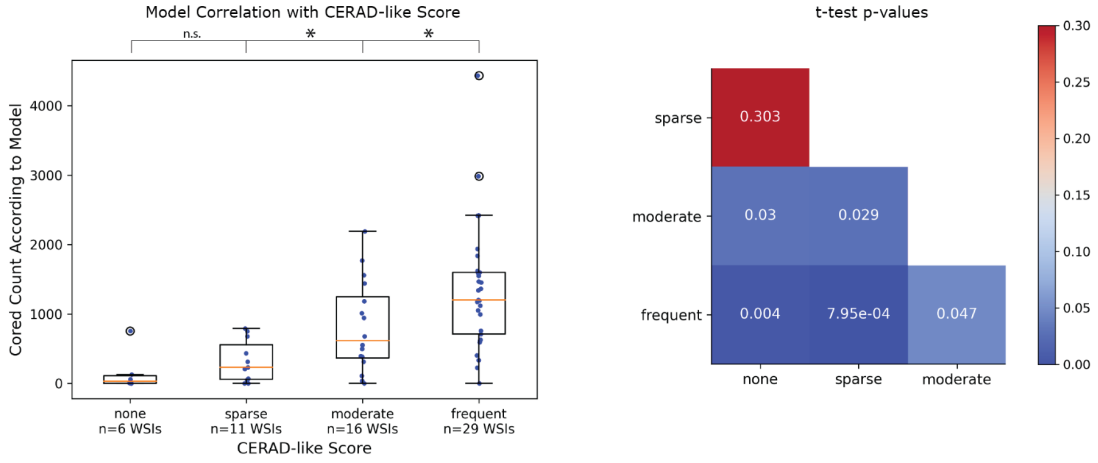


Figure 4



Tables

	YOLOv3 (Titan Xp GPU enabled) 1 min 40 secs / WSI	YOLOv3 (Titan Xp GPU disabled) 5 mins 30 secs / WSI
Tang *et al.*[10] (GTX 1080 GPU enabled) 184 minutes / WSI	110x	33x
Signaevsky *et al.*[16] (Titan Xp GPU enabled) 45 minutes / WSI	27x	8x

Table 1: The model's speed performance compared to other deep learning approaches for quantifying neuropathologies. Speed improvement is the average time of the comparison model divided by the average time of the YOLOv3 model. We note that the other studies used different WSIs and CPUs for benchmarking and that the segmentation model was for tau neuropathologies (although the images are similar in resolution and color space). Tang et al. used a different GPU (Nvidia GTX 1080).

Author Contributions

Conceptualization, D.R.W., B.N.D, M.J.K.; Methodology, D.R.W., M.J.K.; Software, D.R.W.; Validation, D.R.W., H.V.V., W.H.Y, S.D.M, B.N.D.; Formal Analysis, D.R.W.; Investigation, D.R.W., H.V.V., W.H.Y, S.D.M, B.N.D., A.C.M., C.K.W.; Resources, E.S.M., C.D., B.N.D., M.J.K.; Data Curation, H.V.V., W.H.Y, S.D.M, C.K., J.P.G., B.N.D., D.R.W; Writing – Original Draft, D.R.W.; Writing – Review & Editing, D.R.W., B.N.D., M.J.K; Visualization, D.R.W.; Supervision, B.N.D., M.J.K.; Project Administration, D.R.W., M.J.K.; Funding Acquisition, B.N.D., M.J.K.

Acknowledgements

This work was supported by grant number 2018-191905 from the Chan Zuckerberg Initiative DAF, an advised fund of the Silicon Valley Community Foundation (M.J.K.), the National Institute On Aging of the National Institutes of Health (AG 062517 B.N.D.), the University of California Office of the President (MRI-19-599956 B.N.D.), and the California Department of Public Health (CDPH; 19-10611 B.N.D.) with partial funding from the 2019 California Budget Act. We thank UC Davis, UCLA, and UC Irvine under CDPH grant #19-10611 for the validation dataset (released, see Data Availability). We thank Drs. Julia K. Kofler and Charles L. White III for contributions to the training dataset (released in Wong et al.[7]).

Declaration of Interests

The authors declare no competing interests.

Data Availability

All image data is freely available at DOI: 10.17605/OSF.IO/FCPMW (https://doi.org/10.17605/OSF.IO/FCPMW).

Code Availability

The complete source code and fully trained models are available at: https://github.com/keiserlab/amyloid-yolo-paper.

References

1. Shakir, M. N. & Dugger, B. N. Advances in Deep Neuropathological Phenotyping of Alzheimer Disease: Past, Present, and Future. *J. Neuropathol. Exp. Neurol.* **81**, 2–15 (2022).
2. Montine, T. J. *et al.* National Institute on Aging-Alzheimer's Association guidelines for the neuropathologic assessment of Alzheimer's disease: a practical approach. *Acta Neuropathol.* **123**, (2012).
3. Nelson, P. T. *et al.* Clinicopathologic Correlations in a Large Alzheimer Disease Center Autopsy Cohort: Neuritic Plaques and Neurofibrillary Tangles 'Do Count' When Staging Disease Severity. *J. Neuropathol. Exp. Neurol.* **66**, 1136–1146 (2007).
4. Tiraboschi, P., Hansen, L. A., Thal, L. J. & Corey-Bloom, J. The importance of neuritic plaques and tangles to the development and evolution of AD. *Neurology* **62**, 1984–1989 (2004).
5. Fillenbaum, G. G. *et al.* Consortium to Establish a Registry for Alzheimer's Disease (CERAD): the first twenty years. *Alzheimers. Dement.* **4**, 96–109 (2008).
6. Montine, T. J. *et al.* Multisite assessment of NIA-AA guidelines for the neuropathologic evaluation of Alzheimer's disease. *Alzheimers. Dement.* **12**, 164 (2016).
7. Wong, D. R. *et al.* Deep learning from multiple experts improves identification of amyloid neuropathologies. *Acta Neuropathologica Communications* **10**, (2022).
8. Feldman, M. D. Whole slide imaging in pathology: what is holding us back? *Pathol. Lab. Med. Int.* **7**, 35–38 (2015).
9. Hanna, M. G. *et al.* Whole slide imaging equivalency and efficiency study: experience at a large academic center. *Mod. Pathol.* **32**, 916–928 (2019).
10. Tang, Z. *et al.* Interpretable classification of Alzheimer's disease pathologies with a convolutional neural network pipeline. *Nat. Commun.* **10**, 1–14 (2019).
11. Vizcarra, J. C. *et al.* Validation of machine learning models to detect amyloid pathologies across institutions. *Acta Neuropathologica Communications* **8**, 59 (2020).

12. Zhao, Z.-Q., Zheng, P., Xu, S.-T. & Wu, X. Object Detection with Deep Learning: A Review. (2018).
13. Minaee, S. *et al.* Image Segmentation Using Deep Learning: A Survey. (2020).
14. Perosa, V. *et al.* Deep learning assisted quantitative assessment of histopathological markers of Alzheimer's disease and cerebral amyloid angiopathy. *Acta Neuropathologica Communications* **9**, 1–13 (2021).
15. Koga, S., Ikeda, A. & Dickson, D. W. Deep learning-based model for diagnosing Alzheimer's disease and tauopathies. *Neuropathol. Appl. Neurobiol.* (2021) doi:10.1111/nan.12759.
16. Signaevsky, M. *et al.* Artificial intelligence in neuropathology: deep learning-based assessment of tauopathy. *Lab. Invest.* **99**, 1019–1029 (2019).
17. Wang, Y. E., Wei, G.-Y. & Brooks, D. Benchmarking TPU, GPU, and CPU Platforms for Deep Learning. (2019).
18. Liang, X., Nguyen, D. & Jiang, S. B. Generalizability issues with deep learning models in medicine and their potential solutions: illustrated with cone-beam computed tomography (CBCT) to computed tomography (CT) image conversion. *Mach. Learn.: Sci. Technol.* **2**, 015007 (2020).
19. Futoma, J., Simons, M., Panch, T., Doshi-Velez, F. & Celi, L. A. The myth of generalisability in clinical research and machine learning in health care. *Lancet Digit Health* **2**, e489–e492 (2020).
20. Redmon, J. & Farhadi, A. YOLOv3: An Incremental Improvement. (2018).
21. UserBenchmark: Nvidia GTX 1080 vs Titan Xp. https://gpu.userbenchmark.com/Compare/Nvidia-Titan-Xp-vs-Nvidia-GTX-1080/m265423vs3603.
22. McLauchlan, D., Malik, G. A. & Robertson, N. P. Cerebral amyloid angiopathy: subtypes, treatment and role in cognitive impairment. *J. Neurol.* **264**, 2184 (2017).
23. Reinhard, E., Adhikhmin, M., Gooch, B. & Shirley, P. Color transfer between images. *IEEE Comput. Graph. Appl.* **21**, 34–41 (2001).
24. The ultimate training data platform for AI. https://www.superannotate.com/.
25. Everingham, M., Van Gool, L., Williams, C. K. I., Winn, J. & Zisserman, A. The Pascal Visual

Object Classes (VOC) Challenge. *International Journal of Computer Vision* vol. 88 303–338 (2010).