Interpretable classification of Alzheimer's disease pathologies with a convolutional neural network pipeline

Ziqi Tang[1,2], Kangway V. Chuang[1], Charles DeCarli[3], Lee-Way Jin[4], Laurel Beckett[5], Michael J. Keiser[1], and Brittany N. Dugger[4]**

[1] Department of Pharmaceutical Chemistry, Department of Bioengineering and Therapeutic Sciences, Institute for Neurodegenerative Diseases, and Bakar Institute for Computational Health Sciences, University of California, San Francisco, San Francisco, CA, USA

[2] School of Pharmaceutical Sciences, Tsinghua University, Beijing, China

[3] Department of Neurology, University of California - Davis School of Medicine, Davis, CA, USA

[4] Department of Pathology and Laboratory Medicine, University of California - Davis School of Medicine, Davis, CA, USA

[5] Department of Public Health Sciences, University of California - Davis, Davis, CA, USA

*To whom correspondence should be addressed, keiser@keiserlab.org; bndugger@ucdavis.edu

Abstract

Neuropathologists assess vast brain areas to identify diverse and subtly-differentiated morphologies. Standard semi-quantitative scoring approaches, however, are coarse-grained and lack precise neuroanatomic localization. We report a proof-of-concept deep learning pipeline identifying specific neuropathologies—amyloid plaques and cerebral amyloid angiopathy—in immunohistochemically-stained archival slides. Using automated segmentation of stained objects and a cloud-based interface, we annotated >70,000 plaque candidates from 43 whole slide images (WSIs) to train and evaluate convolutional neural networks. Networks achieved strong plaque classification on a 10-WSI hold-out set (0.993 and 0.744 areas under the receiver operating characteristic and precision recall curve, respectively). Prediction confidence maps visualized morphology distributions for WSIs at high resolution. Resulting plaque-burden scores correlated well with established semi-quantitative scores on a 30-WSI blinded hold-out. Finally, saliency mapping demonstrated that networks learned patterns agreeing with accepted pathologic features. This scalable means to augment a neuropathologist's ability may suggest a route to neuropathologic deep phenotyping.

Keywords

Machine learning, Alzheimer's disease, amyloid, deep learning, deep phenotyping, interpretability, explainable AI, convolutional neural networks, dementia, immunohistochemistry, plaques, pathology.

Introduction

Extracellular deposition of amyloid beta (Aβ) plaques is a pathological hallmark of Alzheimer's disease (AD)[1,2], a common chronic neurodegenerative disease. Amyloid plaques have a diverse range of morphologies and neuroanatomic distributions[1]. The current consensus criteria for a neuropathological diagnosis of AD[3–5] incorporate protocols assessing plaque density and distribution; some researchers have hypothesized that plaques may be an initiating event in AD[5,6]. More precise measures of plaque morphologies (such as cored, neuritic, and diffuse) can serve as a basis for understanding disease progression and pathophysiology, providing guidance and insight into disease mechanisms[2,7–10].

For neuropathologic diagnosis, established semi-quantitative scales are used to assess plaque burden (**Fig. 1a**)[4,8,11,12]. The standard semi-quantitative criteria put forth by the Consortium to Establish a Registry for Alzheimer's Disease (CERAD) is based on the manual assessment of the highest density of neocortical neuritic plaques[4,13]. Diffuse plaques, which may be the initial morphological type of Aβ[14,15], can account for over 50% of plaque burden in preclinical cases but are not included in CERAD[16]. Furthermore, data on anatomical location (*i.e.* Thal amyloid phase) are based on the presence of plaques regardless of type or density[5]. The potential for neuropathologic deep phenotyping efforts that account for anatomic location, diverse sources of proteinopathy, and quantitative pathology densities motivates the development of effective and scalable quantitative methods to differentiate pathological subtypes[17–19].

Existing quantitative methods, such as positive pixel count[20] algorithms, rely on human-defined[21] red-green-blue or hue-saturation-value ranges (*i.e.*, pixel color and intensity) and are thus sensitive to batch differences or to the variable effects of formalin fixation on tinctorial properties. Manual counts or stereological[22,23] methods can be tedious, difficult to score, and time-consuming. Consequently, studies using limited-range scores or overall pathology burden[3,8,13,24] are powerful but have interrater variability,[4,13] are difficult to adapt to statistically meaningful disease-correlation analysis, or are blind to selected locational vulnerability[20]. New methods introducing detailed and sensitive quantification of pathologies would reduce the burden placed on pathologists, increase reliability, and enable studies at a scale that is currently prohibitive.

Deep learning has transformed medical image analysis[25,26]. Convolutional neural networks (CNNs) have achieved expert-level performance in complex visual recognition tasks, including the diagnosis of skin[27] and breast[28,29] cancers. These flexible models learn to recognize intricate patterns directly from visual data without the need for manually-defined image features or expert-delineated templates, and can account for non-trivial variations in image quality and color. In neuropathology, deep learning approaches have been reported for the classification of AD pathophysiology in magnetic resonance and positron-emission tomography images[30], and for relating gene expression to neuropathology datasets[31].

We hypothesized deep learning methods could augment neuropathological whole slide image (WSI) analysis[32]. Despite their strong predictive power, deep learning models have been criticized for their poor interpretability and reliance on massive annotated datasets[33]. At the outset, we recognized these factors represented significant challenges in the development of useful tools for neuropathology. An approach tailored to neuropathology would require 1) careful delineation of the machine learning task; 2) construction of a curated image dataset with high-resolution annotations by experts; and 3) extensive model interpretability. As a proof of concept, we posited CNN models could be employed for recognition and classification of Aβ pathologies, especially plaques, with the downstream goal of providing reliable, scalable, and interpretable measures based on neuroanatomical location. To develop a useful tool to aid neuropathologists, we deemed it critical that predictive performance should result from learning meaningful patterns within the images[34].

In this study, we present a pipeline for the neuropathological analysis of Aβ pathologies in WSIs generated by digitizing glass microscope slides of temporal gyri of human brain (**Fig. 1b**). We describe an end-to-end pipeline for image processing, a custom web interface for rapid expert annotation, and training of CNN models that result in high performance multi-task classifiers capable of distinguishing Aβ in the form of cored plaques, diffuse plaques, and cerebral amyloid angiopathy (CAA). We demonstrate how prediction confidence maps visualize distributions as an interpretable and complementary means to understand Aβ burden. Finally, we provide visual evidence that these models are interpretable, using deep learning introspection methods to show that trained models learn relevant features of each of these Aβ pathology classes. To the best of our knowledge, these studies constitute the first report of CNNs for Aβ pathology analysis.

Results

A Platform to Rapidly Manually Annotate 77,000 Plaque Candidates

CNNs operate most effectively when trained on datasets exceeding tens of thousands of example images[35]. Indeed, we found that 43 digitized glass microscope slides (WSIs, see **Supplementary Table 1** for case details) used in this study yielded over 500,000 individual candidate plaques. We set out to build an dataset of approximately 50,000 annotated images for model training (**Fig. 2**, see **Data Availability**). Manual annotation at this scale would have been a daunting task through conventional, hand-drawn bounding boxes on a standard ~700 micron visual field. Using open-source image analysis tools (see **Methods**), we developed an automated preprocessing procedure (**Fig. 2a** and **Supplementary Fig. 1, 2**) to normalize slide color and generate bounding boxes around all immunohistochemically-stained objects within WSIs. As the native resolution of a WSI is too large (typically 50,000 by 50,000 pixels at 20x magnification) to use as the direct input for CNNs, we designed the dataset to contain uniform 256x256 pixel tiles centered on individual plaque candidates.

We created a simple web interface to rapidly annotate Aβ pathology-candidate image tiles and deployed it on the Amazon Web Services Elastic Beanstalk[36] for reliability and scalability (illustrated in **Fig. 2c**). An expert neuropathologist annotator used unique credentials and a rapid keystroke-entry format to annotate the tiles, which were stored in a SQL database (see **Supplementary Fig. 3**). Using this platform, candidate images were annotated at rates up to 2,500 tiles per hour into three major categories—“cored” plaques, “diffuse” plaques, or cerebral amyloid angiopathy (“CAA”). Additional categories such as “not sure” or “flag” denoted uncertainty, image segmentation failures, or other special cases (**Supplementary Fig. 3**). The dataset was then built in three phases (**Table 1**)**.** In Phase I, 55,001 images were expert-labeled using the web application. The majority of candidate images were annotated as diffuse plaque morphologies (84.8% of the annotations), with cored plaques (2.2%) and CAAs (1.1%) making up the minor classes. As class balance typically improves machine learning model performance, we sought to enrich the minority classes in a second phase. We trained an intermediate CNN to classify objects based on the Phase I dataset, then used its predictions to prioritize an additional set of 101,671 unprocessed tiles in favor of cored plaques and CAAs for manual annotation (see **Supplementary Fig. 4**). Thus in Phase II, an additional 11,029 tiles were annotated, having been evaluated in rank-order of their predicted likelihood to contain either of the minority-class plaques. In Phase III, we annotated an additional 10,873 plaque candidate tiles extracted from a separate hold-out test set of 10 WSIs not in the original 33-WSI collection, without any prioritization procedures. We performed manual annotation using the web application for all phases.

Convolutional Neural Networks Effectively Discriminate Among β-Amyloid Morphologies

We trained CNNs to classify tiles as containing cored plaque, diffuse plaque, or CAA. At 20x magnification, a single 256x256 pixel tile (128 microns) could contain more than one object, so we trained a multi-task CNNs for multi-label classification: CNNs were asked to determine the presence or absence of all morphologies in each tile. We combined the Phase I and Phase II datasets, then randomly split the resulting 70,000 tiles (66,030 annotated and 3,970 IHC-negative) into *training* (from 29 WSIs) and *validation* (from 4 WSIs) sets, while stratifying by case (*i.e.*, WSI source) to ensure that models generalize to new cases. A search of CNN architectures identified a six-layer convolutional architecture with two dense layers (**Fig. 3a**) with strong performance. Using subsequent hyperparameter optimization we found that data augmentation[37,38] and minority class oversampling[39] training procedures (**Supplementary Fig. 5**, **Methods**) yielded a pronounced performance boost.

The resulting CNN model trained on 61,370 example tiles achieved validation set performance of 0.983 area under the receiver operator characteristic (AUROC) (**Supplementary Fig. 6a**) with an area under the precision recall curve (AUPRC) of 0.842 (**Supplementary Fig. 6b**). On the strict hold-out (Phase III) test set, the model likewise generalized well to unseen decedent cases (AUROC = 0.993, AUPRC = 0.744, **Fig. 4a**, **4b**). CAA prediction performance was also strong on the validation set (**Supplementary Fig. 6**), but was omitted from the Phase-III test set benchmarking in **Fig. 4** because Phase III was derived from neuroanatomic regions predominantly lacking in CAAs (e.g., **Supplementary Fig. 7**). Notably, model performance was achieved using fewer than 2,000 training examples of each minority class (cored plaques and CAAs). Representative accurate (**Fig. 3b**) and misclassified (**Fig. 3c**) examples from the 10,873 hold-out set tests illustrate cases where the model succeeded or went astray.

Model Performance Improves Nonlinearly with the Number of Training Examples

To determine whether similar performance could be achieved with fewer manual annotations, we performed two retrospective studies to investigate the effect of training dataset size. In the first study, we randomly selected subsets of the 61,370-example training dataset, maintaining stratification by case (*i.e.*, WSI source), and plotted model performance as a function of the number of training examples (**Fig. 4c**). Each random selection was repeated five times, and a fresh model trained each time, for a total of 95 independently trained and evaluated CNN models with identical architectures. All models were benchmarked against the same hold-out (Phase III, as in **Fig. 4a-b**) test set. As expected, model performance positively tracked with the total number of training examples. Notably, models trained on a 50% smaller training set size still achieved an average AUROC above 0.99 and an AUPRC above 0.74, at minimal loss to overall performance.

In the second study, we investigated model performance as a function of the chronological dataset growth during the project, where training examples were included in the order of original expert annotation (**Fig. 4d**). Model performance at 15 expert-hours fell short of model performance at 50% of dataset size (**Fig. 4c**). Accordingly, the goal of this second study was to determine whether annotation chronology played a role in CNN training. As above, performance steadily increases as the annotated dataset grows. However, performance trends between the studies differed in two ways. Chronologically-trained models did not converge in AUPRC performance as early as the equivalent-sized random-subset-trained models benefitting from later annotations did. Secondly, the chronology study shows a distinct AUPRC boost in Phase II, illustrating the positive effect of enriching for cored-plaque prevalence.

Prediction Confidence Maps Show Plaque Localization

To visualize the distribution and neuroanatomic location of pathologies in a broader context, we applied a sliding window approach[40] to generate WSI heatmaps of predictions (**Fig. 5**). These heatmaps plot the confidence and location of each prediction by the CNN, which may then be visualized from the sub-tile resolution (**Fig. 5c**) up to the full WSI view (**Fig. 5a**). By progressively zooming in from larger anatomical views, the visualization shifts from the broad distribution of plaques to their detailed 20x morphology. A single cored plaque can be distinguished from a dense region of neighboring diffuse plaques (**Fig. 5c**). In this cohort, diffuse plaques are densely distributed across the grey matter, whereas cored plaques are located in deeper and lower cortical layers, in accordance with known neuroanatomic distributions[1]. Furthermore, CAA predictions appear predominantly at the periphery, proximal to the cortical surface where leptomeninges are present[41], although predictions are made independently of the surrounding field or broader neuroanatomic context. These maps highlight other locational aspects of the plaques, such as their presence in white matter immediately beneath the gray matter[42].

Classification Performance Does Not Vary by Tissue Landmarks

The CNNs perform classification (e.g., **Fig. 4**) directly on small anatomical areas (128 microns; green box in **Fig. 6a**). Human experts typically assess larger fields of view such as ~700 microns viewed at 10x magnification when conducting semi-quantitative plaque scoring. To visualize prediction performance in this context, we also assessed cored-plaque agreement maps on contiguous 6-by-6 tile (768 micron) regions (**Fig. 6a**). In the leftmost column, a green box surrounds the cored plaque within the tile, as labeled by a neuropathologist during the Phase-III dataset annotation (**Fig. 6a**). The middle column overlays the prediction map (as in **Fig. 5c**) onto the original IHC-stained image. Finally, the rightmost column summarizes agreement between the expert label and the prediction, with blue and cyan representing correct prediction areas, while red and orange denote misclassification[43]. For this analysis, we used a CNN prediction confidence threshold of 0.90. A more permissive threshold would decrease false negatives (red) at the cost of more false positives (orange). Interestingly, this agreement-map highlights the limitations of bounding-box annotations, such that the correct cored-plaque prediction shown is nonetheless penalized by this view (red halo) for accurately predicting the rounded boundaries of the actual plaque instead of anticipating its square “ground truth” bounding-box.

Taking a step yet further out to regions of 3840 microns (**Fig. 6b**), these maps (see **Supplementary Fig. 8** for additional examples) visualize the results plotted in **Fig. 4**, with the complementary addition of tissue landmarks, prediction clustering, and neuroanatomic localization. The model reliably rejects background tissue and diffuse plaque deposits, while accurately identifying most cored plaques. Model performance does not change based on nearby neuroanatomic architecture in these examples, although occasional clusters of co-localized false-positive (orange) cored plaque predictions can appear (e.g., **Supplementary Fig. 8**).

Machine Learning Introspection Techniques Identify Salient Plaque Features

To investigate the CNN model’s internal logic, we performed two studies to determine the importance of morphology features contributing to accurate predictions (**Fig. 7**). In the first, we applied guided gradient-weighted class activation mapping (guided Grad-CAM)[44] to identify salient visual features underlying the model’s predictions. Guided Grad-CAM follows the CNN’s gradient flow from individual tasks back onto the original image tile to establish an activation map, highlighting the input features most relevant to each CNN prediction. **Fig. 7** shows guided Grad-CAM (white features on black background) on examples taken from cored, diffuse, and CAA plaque classes. Consistent with human expertise, Grad-CAM activation maps predominantly highlight regions of the tiles corresponding to IHC-stained amyloid deposits. For instance, in **Fig. 7a,** the activation map for cored-plaque prediction highlights a dense and compact amyloid center (yellow arrow), whereas for the diffuse task the activation map highlights the off-center diffuse object (red arrow). By contrast, the CAA activation map highlights the periphery of the image, much as CAA often forms a ring within vessels, although none could be found. In **Fig. 7b**, the diffuse-task Grad-CAM highlights ill-defined amorphous amyloid deposits, while for the cored and CAA tasks it focuses on punctate IHC staining and microglia. The

CAA activation map in **Fig. 7c** highlights ring structures (blue arrow) within the media of the cortical vessel, consistent with CAA's defining feature; while for cored and diffuse tasks, Grad-CAM highlights the punctate deposit (red arrow) beneath the CAA. Lastly **Fig. 7d**, which contains both a diffuse (red arrow) and a cored plaque (yellow arrow), shows cored-task activation maps localizing to the amyloid core, with broader feature activations for diffuse and CAA tasks. Crucially, Grad-CAM activation mapping may highlight certain image features as salient because they help determine that an object is *not* present in the image: Despite strong localized activation for cored and diffuse maps in **Fig. 7c** at the punctate deposit (red arrow), the CNN predicts that neither plaque is present.

Whereas guided Grad-CAM provides a fine-grained view of feature salience, it does not differentiate features indicative of a plaque from those that contradict its presence. To complement the analysis, we performed a feature occlusion study[45] on the same examples. In this experiment, a small occlusion patch (shown in **Fig. 7a**, black box) is systematically moved across the image, and the model makes a prediction on the occluded image at each increment. Blue-to-yellow-to-red colors indicate increasing CNN prediction confidence from 0.0 to 1.0. Consequently, color shifts in occlusion maps show which image features, when occluded, change prediction confidence. When the patch occludes an important feature such as the amyloid core of a cored plaque (**Fig. 7a**, yellow arrow), the model fails to predict the object correctly: cored-task confidence drops to zero (blue dot on red background, yellow arrow). Occluding less cored-task-relevant relevant regions such as within the off-center diffuse stain (red arrow) have little effect, indicated by the solid red coloring in the cored-task's confidence map for this area. Conversely, confidence maps may also show where occlusion of a critical feature makes an alternative class more likely. If the amyloid core in **Fig. 7a** is occluded, diffuse plaque prediction becomes likely (signified by yellow arrow).

Where more than one plaque occur within the same tile, the two feature-importance studies differ. Grad-CAM activation maps identify salient pixels for plaque classes independently, whereas occlusion maps highlight the interplay of features among classes. For example, in occlusion maps, occluding the leftmost plaque decreases diffuse-task confidence (**Fig. 7d**, light blue region in diffuse-task map, red arrow), whereas the same prediction gains confidence (red region, yellow arrow) when the cored plaque's core is occluded. In the corresponding grad-CAM activation maps for the diffuse-task, however, features specific to the diffuse plaque are predominant. Together, these complementary maps visualize the features within images that motivate the CNN's plaque predictions.

Model-Based Whole-Slide Scores Correlate with Manual Semi-Quantitative Scores

To compare with manual semi-quantitative approaches such as CERAD, we developed a preliminary neural-network derived score for plaque burden at a global WSI level. For the CNN-based score, we calculated a count of predicted plaques across an entire WSI by segmenting its prediction heatmap (e.g., **Fig. 5a**) and normalizing the result by the tissue area of each slide (**Supplementary Fig. 9**). The resulting CNN-based scores correlated strongly across the total dataset of 63 WSIs (**Supplementary Tables 1** and **2**) for which we had independently-collected semi-quantitative, CERAD-like scores for each specific class (**Fig. 8**). CNN-based scores for Aβ significantly differentiated WSIs by CERAD-like categories (e.g., “moderate” vs. “frequent”), especially for cored plaques (**Fig. 8a**, second row). For instance, CNN-based WSI scores between “none” versus “frequent” CERAD-like categories were exponentially separated. To better assess generalization, we collected a further set of 20 WSIs **(Supplementary Table 2**) with corresponding CERAD scores that were blinded during analysis. Combined with the 10 separate hold-out WSIs from Phase III, we found this 30-WSI blinded hold-out set demonstrated strong correlation between the automated and manual scoring approaches, such that CNN-based scores significantly discriminated existing semi-quantitative categories (**Fig. 8b**).

Discussion

We report a scalable, quantitative, and interpretable approach to identify neuropathologies for three classes of Aβ deposits, motivated by the method’s downstream application to statistically powerful correlative analyses and neuroanatomical localization of AD pathologies. In practice, such deep-phenotyping techniques will have limited utility if their underlying predictions cannot be interpreted, critiqued, and refined by expert neuropathologist supervision. Consequently, to establish the feasibility and limitations of this approach, we considered multiple challenges when adapting CNNs to WSIs of archival human brain samples[46]: 1) WSIs are prohibitively large and can vary in color, necessitating a preprocessing pipeline for their normalization, segmentation, and efficient annotation; 2) deep neural networks are notoriously data-hungry; we evaluated how much training data a CNN requires to discriminate Aβ morphologies (cored plaque, diffuse plaque and CAA); and 3) it is rarely obvious how generalizable a model will be across diverse individuals, populations, and conditions—to prove an enabling tool, model predictions must be interpretable and derive from meaningful features amenable to a neuropathologist’s critique and feedback.

Addressing the first challenge, we developed an end-to-end pipeline to automate WSI processing and aid rapid image annotation (**Fig. 2a**, **Supplementary Fig. 3**). This pipeline performs color normalization[47] followed by immunohistochemical (IHC) stain detection to create a preliminary library of candidate plaques at 20x magnification. The logic behind stain detection was two-part: Stained objects inhabit a delimited brown hue range and objects comprise coherent contiguous regions exceeding a minimum size. We then generated image tiles centered on each candidate (**Fig. 2b**). Thus, 43 WSIs containing Aβ IHC-stained temporal gyri yielded nearly 500,000 raw candidate tiles at 20x resolution (details in **Supplementary Table 1-2**), filtered to approximately 200,000 tiles containing candidate objects of sufficient size (see **Methods**). The next step was image annotation for supervised machine learning. Although web-based histopathological annotation tools exist[48,49], we developed a simple platform using the cloud-based Amazon Web Services Elastic Beanstalk[36] infrastructure (**Supplementary Fig. 3**) for study design flexibility and for the speed of its keystroke-based entry format. For instance, subsequent studies may investigate a broader field of view for annotation context or introduce checks on intra-rater reliability by re-presenting tiles to annotators in different orientations. Given the scope of the annotation task, we incorporated several aspects of gamification theory[50,51], such as annotator leveling, achievement badges, and progress-bar filling[52,53] to acknowledge and motivate progress. Using this tool, we observed sustained annotation rates reaching 1.44 seconds/tile.

The second challenge was in determining the necessary training dataset size. Having manually annotated 66,030 candidate tiles from 33 WSIs in two annotation phases (**Table 1**), plus 3,970 randomly selected IHC-negative tiles, we examined the CNN's ability to precisely discriminate plaque and CAA morphologies. For this analysis, we randomly split the tiles into train and validation sets, such that train and validation tiles never shared the same WSI source. Additionally, as a strict hold-out test set (Phase III) and to investigate the role of nearby neuroanatomic landmarks on prediction, we annotated larger contiguous tissue regions corresponding to 5x the standard dimensions, for 10 previously unseen WSIs (**Supplementary Fig. 7**). Note this Phase III dataset differed from the Phase I+II train and validation datasets in that the latter's 70,000 labeled tiles were randomly selected; there was no guarantee that tiles and their resulting expert plaque annotations would be contiguous or comprehensively labeled for any local region of tissue. We trained multi-task CNNs on all 61,370 training tiles, evaluated multiple CNN architectures and hyperparameter choices, and found that a relatively simple model design (**Fig. 3a**) inspired by Simonyan and Zisserman's VGG[54] achieved strong classification performance (**Fig. 4a-b**).

Given the substantial time investment, we asked whether similar performance could have been achieved with fewer training tiles. We evaluated this retrospectively, by progressively decreasing the training dataset size in two different ways (**Fig. 4c-d**). In the first study, we selected a progression of training data subsets randomly and repeated the training process five times per subset size (**Fig. 4c**). In the second study, we maintained the chronology of the project instead, and plot a natural history of the annotation process. Intriguingly, these performance evaluations highlighted two annotation regimes (**Fig. 4d**); first, unbiased random-tile candidate labeling (Phase I), followed by the Phase II procedure, where cored-plaque and CAA candidates were purposefully enriched by bootstrapping from a Phase-I-trained CNN model. As expected, increasing training example counts improved model performance. Less anticipated was that chronologically-early annotations appeared to be less effective for model training (**Fig. 4d**); considerations such as the neuropathologist's growing familiarity with the annotation tool and its visual field may be subjects for further study. From a practical perspective, the steepest performance gains were achieved within the first 15 hours of expert labeling, suggesting a reduced dataset may be pragmatically sufficient for classification of cored and diffuse plaques. Significantly, models trained using a comparatively small investment of a neuropathologist's time can assist with new cases and potentially reduce overall expert burden. Subsequent refinements to the model, particularly in reinforcement feedback on incorrectly-classified examples encountered during the model's use (e.g., **Fig. 3d**), might later be incorporated into the workflow with minimal friction.

The third challenge was human interpretability. We posited that visualizing the CNN model's predictions as comprehensive confidence maps from the whole-slide level down to a focused plaque-level field (20x) would aid interpretability by a trained neuropathologist, given the importance of local tissue and neuroanatomic context. On a neuroanatomic level, most predicted plaques are located within grey matter (**Fig. 5a**, yellow-to-green regions, right three columns) with some sparse densities in white matter not appreciated from the raw slide (**Fig. 5a**, left column). Despite their primary localization within grey matter, studies have reported plaques within white matter[1,42,55]. Furthermore, the maps predict cored plaques' propensity for deeper and lower cortical layers, consistent with their known neuroanatomic distribution[1,56]. We were likewise gratified to observe that individual cored plaques stand out from dense neighborhoods of diffuse plaques (**Fig. 5c**, cored column) and that CAA predictions made by the model on a 20x (128 microns) tile-by-tile basis nevertheless localized predominantly to the leptomeninges and some within cortex grey matter. There were caveats however, as for instance when clusters of diffuse plaques having staining "halos" were misclassified as CAAs (**Fig. 5b**, CAA column). This was not entirely surprising as the project focused on cored plaques, so the CAA dataset was comparatively small; a larger CAA dataset containing the full spectrum of its morphologies may be a useful subject of further projects. Indeed, CAAs can be delineated into various staging schemes, such as by their location within the media of the vessel and vessel integrity[11], which is important in diagnosing CAA-related hemorrhage.

Using Phase III's larger field-of-view (3840 microns) hold-out regions, the overlays of CNN prediction confidence maps onto ground-truth annotations highlighted cases and context of prediction success and disagreement (**Fig. 6b**, **Supplementary Fig. 8**). Nonetheless, accurate predictions alone do not guarantee meaningful learning or that the model will be applicable to new scenarios or populations[57]. Plaque morphology can differ by neuroanatomic location—a CNN model developed from temporal gyri plaques may not be translatable to plaques in other anatomic areas, such as the striatum[1,58]. Although an explicit evaluation of all confounders is outside the scope of this work, the feature saliency and occlusion map studies (**Fig. 7**) demonstrated that the models focus on image features relevant for neuropathology. Guided gradient-weighted class activation mapping (guided Grad-CAM) techniques near-exclusively highlighted the IHC stained regions, in patterns characteristic of the pathologies (**Fig. 7**, white-on-black maps). Complementarily, feature occlusion studies illustrated that the central amyloid core is the most discerning feature of a cored plaque's correct identification, and that its occlusion transforms a CNN model's classification to diffuse plaque. Importantly, the crucial features emerging from these machine learning introspection techniques—dense compact amyloid centers for cored plaques, ill-defined amorphous amyloid deposits for diffuse plaques, and amyloid within the media of the cortical vessels for CAAs—all agree with key features used by experts[1,11,59,60].

We finally evaluated whether CNN models could automatically quantify plaque burden on a whole-slide level in a way that would correlate with standard semi-quantitative methods for plaque assessment (i.e. CERAD neuritic plaque scores). As true neuritic plaques are not distinguishable using Aβ-selective IHC stains, we leveraged CERAD-like manual scores (none, sparse, moderate and frequent) specific to each amyloid class. We found that a preliminary WSI-level CNN-based score we developed (**Methods**) correlated strongly with manual CERAD-like scores (**Fig. 8**). CNN-based scores from one CERAD-like category were significantly different from WSIs in other categories (for cored plaque, $p<0.01$). Beyond its overall correspondence with

CERAD, the finer-grained CNN-based metric captured subtle variations of plaque burden within each CERAD category. The more detailed and sensitive measurement of plaque burden, after appropriate validation in further studies, may strengthen statistical power for clinicopathological correlations[61]. Automated scores of this nature might be applied across entire archives of stained tissue from diverse anatomic regions, or aid in studies focused on evaluating burden specific to certain neuroanatomic locales or other local landmarks.

Several caveats, however, merit mention. Foremost among them is the intentional restriction of this proof-of-concept study's scope to annotations made by a single expert neuropathologist on a single immunohistochemical stain within a single anatomic region. Differences in experience and annotation criteria will likely result in individual expert variation among "ground truth" labels. The goal and intent of this project was therefore to establish the potential to extend an individual neuropathologist's plaque-identification capabilities in the context of their normal workflow. Furthermore, all data used in this study were from a single brain bank and retrieved and digitized under the same conditions; more diverse datasets from multiple sources will yield more robust and reliable models. We noted also that when the same hold-out set (Phase III) was annotated by web platform (**Fig. 2c**) versus entirely by hand, this resulted in at times differing labels (**Supplementary Fig. 10**). Future work may build on these foundations to investigate cross-neuropathologist plaque labeling, differing stains, anatomic regions, or collection centers, as well as region-level scoring systems to quantify bulk Aβ deposit burdens.

Conclusion

Taken together, the present study demonstrates a deep learning approach that can augment the expertise and analysis of an expert neuropathologist. Approximately 30 hours of expert labeling yield a highly scalable and reusable CNN model capable of novel inference on unseen WSIs—going forward, 15 hours may be sufficient to create new visualizations and understandings of plaque distribution. CNNs automatically learn relevant features from immunohistochemical image data and to exploit Alzheimer's disease pathological features in agreement with human-interpretable neuropathology. Significantly, CNN-based Aβ pathology burden agrees with CERAD-like scoring. Many brain banks have archival materials of stained slides arising from neuropathological diagnosis; models such as these may capitalize on such materials to help quantify pathologies providing pathological deep phenotyping in a scalable way. We anticipate that collecting annotated datasets from multiple sources and experts will improve recall, sensitivity, and accuracy of the resulting neural network models and support training of more sophisticated model architectures. We hope that this proof of concept motivates further work in this field, where automated pathology classification could have far-reaching impact; to this end, we make the CNN model code and dataset openly available to the community (see **Data Availability**).

Materials and Methods

Dataset

Case Cohort – All samples were retrieved from archives of the University of California, Davis Alzheimer's Disease Center (UCD-ADC) Brain Bank. Archival samples analyzed in this study were 5 µm formalin fixed, paraffin embedded, sections of the superior and middle temporal gyrus. The tissue had been previously stained with an amyloid-β antibody (4G8, recognizing residues 17-24, BioLegend, formerly Covance) that were first pretreated with formic acid to rid samples of endogenous protein. All slides were digitized using an Aperio AT2 up to 40x magnification.

Dataset Splitting – A total of 33 WSIs corresponding to 33 separate decedent cases, spanning non-AD to AD and possessing a variety of CERAD scores (see **Supplementary Table 1**) were used for model development and training (29 training, 4 validation images). An additional 10 WSIs were selected by an expert neuropathologist (BD) as a held-out test set and were not released until the model development phase of the study had been completed. Finally, a further 20 blinded WSIs were collected solely for use in the CERAD-like scoring comparison study (**Fig. 8b**). CERAD-like scores were available for all but one of the 63 WSIs used in this study; thus **Fig. 8a** uses 62 WSIs as a complete dataset.

Image Preprocessing – All initial image preprocessing was performed in the open-source library PyVips[62]. Images were loaded at 20x magnification, corresponding to a resolution of 0.5 microns per pixel (MPP). Slide color normalization was performed by the method of Reinhard et. al[63] using a reference image selected by the annotating expert (see **Supplementary Fig. 1**). The resulting WSIs were regularly tiled to 1536x1536 pixel tiles, corresponding to 768x768 micron regions of tissue for further analysis, resulting in a total of 33,111 tiles for the training set.

Segmentation – Image segmentation was performed using the open-source library OpenCV[64]. Immunohistochemically-stained entities including cored plaques, diffuse plaques, and CAAs appeared in the brown hue-region and segmentation was performed in the HSV colorspace utilizing a permissive colormask. Morphological opening and closing operations were performed to smooth the binary masks, and a standard "blob"-detection procedure was applied to isolate candidate objects. These unique components were center-cropped to a fixed size (256x256 pixels), corresponding to a region of 128x128 microns. This procedure resulted in nearly 500,000 images. Noisy background deposits were eliminated through a minimum size cutoff, resulting in a total of 206,888 tiles.

Plaque-labeling Web Interface

To allow for the rapid and efficient annotation of the dataset, we developed a custom Python Flask web application that we deployed on Amazon Web Services Elastic Beanstalk[36]. The web-based interface allowed for remote login by the expert labeler, and enables fast, multi-label annotation of images using individual keystrokes. In the interface, images corresponding to 128x128 microns regions were shown to the annotator. A bounding box in the image specified which specific plaque candidate was to be labeled. Several elements of gamification, such as leveling, achievement badges (crown icon), and progress bar filling (green bar) were incorporated to motivate and track annotation task progress. A timestamp function was implemented to record the number of images/hour annotated by the expert (BD). All labels were stored in a SQL database using the Amazon Relational Database Service.

All images were annotated by a single neuropathologist (BD) and labeling of the image data proceeded in three phases: 1) In an initial phase, 55,000 images stemming from 3,811 unique tiles were labeled; 2) In the second phase, images containing the minor classes of interest (cored plaques and CAAs) were enriched by running the CNN model built from the first-phase dataset on the remaining 101,671 images. These images were ranked by their predicted likelihood of containing cored plaques or CAAs. We then chose the top 11,029 images for labeling. The labeled data from Phase I and Phase II were combined as the entire dataset (Phase I+II) for model training and evaluation. 3) In the third phase, two test sets were constructed with the same data but two distinct labeling methods. A 7680 x 7680 pixel (0.5 MPP) region was selected within each of the 10 hold-out test set WSIs by an expert neuropathologist as the area of interest. For the first test set, 10,873 plaque candidate tiles extracted from these 10 regions were labeled using the plaque-labeling web interface. For the second test set, the cored plaques and CAAs were directly marked by a neuropathologist on the selected region at a standard 10x (768 microns) visual field.

Model Development and Training

CNN Model Architecture and Training – All neural network models were trained in the open-source package PyTorch[65] on four NVIDIA GTX 1080 or Titan X graphics processing units. Our optimized model used a simple convolutional architecture for image classification, consisting of alternating (3x3) kernels of stride 1 and padding 1 followed by max pooling (**Fig. 3a**), followed by two fully connected hidden layers (512 and 100 neurons) and rectified linear units as the nonlinear activation function. All neural network models were trained using backpropagation. The optimized training procedure used the Adam[66] optimizer with a multi-label soft margin loss function with weight decay (L2 penalty, 0.008) and dropout (probability 0.5 for the first two fully connected layers and probability 0.2 for all convolutional layers). Training proceeded with mini-batches of 64 images with real-time data augmentation including random flips, rotations, zoom, shear, and color jitter.

Prediction Confidence Heatmaps

A sliding window approach[40] was applied with the trained CNN model to generate confidence heatmaps. At each step, the CNN model took a 256x256 pixel region as input and generated a prediction score for cored plaques, diffuse plaques and CAAs. By systematically sliding the input region across the entire image, the prediction scores were plotted as prediction confidence heatmaps. The color represented the CNN's prediction confidence for the presence of cored plaques, diffuse plaques, and CAAs in the corresponding region, with yellow being the most confidence, and purple the least. We used a stride of 16, 4, and 1 for **Fig. 5a-c** respectively.

Guided Gradient-Weight Class Activation Maps (Guided Grad-CAM)

Guided gradient-weighted class activation mapping (guided Grad-CAM)[44] was performed in PyTorch to generate saliency maps that highlight relevant features on the class of interest. The saliency map is a pointwise multiplication of guided backpropagation and grad-CAM. Guided backpropagation produces a pixel-space gradient map of predicted class scores with respect to pixel intensities of input image. Grad-CAM produces a more class specific map which is the dot product of the feature map of the last convolutional layer and the partial derivatives of predicted class scores with respect to the neurons in the last convolutional layer. We employed an open-source implementation of Grad-CAM by Ozbulak[67] along with our trained models to evaluate the learned features of the model on individual class examples (**Fig. 7**).

Feature Occlusion Studies

Feature occlusion studies[45] were performed to show the influence of occluding regions of input image to the confidence score predicted by the CNN model. The occlusion map was computed by replacing a 16x16 pixels region of the image with a pure white patch and generating a prediction on the occluded image. As systematically sliding the white patch across the whole image (stride = 1 pixel), the prediction score on the occluded image was recorded as an individual pixel of the corresponding occlusion map. The color represented the CNN's prediction confidence for the presence of cored plaques, diffuse plaques, and CAAs of the occluded image, with red being the most confidence, and blue the least.

CNN Model WSI Scores

CNN model WSI scores were generated through segmentation of WSI prediction heatmaps (see Prediction Confidence Heatmaps, above) and normalizing the predicted count by the WSI tissue area.

Tissue Segmentation – Tissue areas from WSIs were calculated using the open-source libraries PyVips and OpenCV. Tissue segmentation against the slide background was performed by applying a color mask in the lightness-chroma-hue (LCH) colorspace. Morphological opening and closing operations were performed to smooth the binary mask, and the tissue areas were calculated as the pixel sum of the refined mask.

Segmentation on Prediction Confidence Heatmaps – Prediction confidence heatmaps were segmented using the open-source library OpenCV[64]. First, a CNN confidence threshold was applied to the heatmaps, with only prediction confidences higher than the threshold retained, indicating the positive predictions of plaques. Morphological opening and closing operations were then performed to smooth the binary masks, and prediction areas exceeding a second threshold set to eliminate CNN-noise provided predicted quantities of the Aβ pathologies by the CNN model. These hyperparameters, CNN confidence threshold (cored plaque: 0.1; diffuse plaque: 0.95; CAA: 0.9) and size threshold (cored plaque: 100; diffuse plaque: 1; CAA: 200), were optimized by statistical analysis on the training and validation sets.

Statistical Analyses – Statistical analyses were performed using the open-source library SciPy[68]. A two-sided, independent, two-sample *t*-test was used to test the null hypothesis that two independent samples have identical expected values. CNN scores of WSIs from different CERAD categories were used for the test. Data were presented as box plots overlaid with individual data points. Box plots showed interquartile range (top and bottom of the box), median (central band), and outliers (points beyond the whiskers). Individual data points were shown as specific points. $P \geq 0.05$ was considered not significant (ns); *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$.

Data Availability

We have made the CNN model code (https://github.com/keiserlab/plaquebox-paper/) and the full raw WSI dataset and the annotated plaque-level dataset (https://www.keiserlab.org/resources/) openly available.

List of Abbreviations

Aβ - Amyloid-beta

AD - Alzheimer's disease

AUPRC - area under the precision recall curve

AUROC - area under the receiver operator characteristic

CERAD - Consortium to Establish a Registry for Alzheimer's Disease

CNN - convolutional neural network

CAA - cerebral amyloid angiopathy

FN - false negative

FP - false positive

grad-CAM - gradient-weighted class activation mapping

HSV - hue-saturation-value

IHC - immunohistochemical

LHC - lightness-chroma-hue

MPP - microns per pixel

RGB - red-green-blue

SQL - standardized query language

TP - true positive

TN - true negative

WSI - whole slide image

Declarations

Ethics approval and consent to participate

These studies utilized only human post-mortem tissues. Only living subjects are defined as Human Subjects under federal law (45 CFR 46, Protection of Human Subjects). All participants or legal representative signed informed consent during the life of the participant as part of the University of California Davis

Alzheimer's Disease Center program. All human subject involvement was overseen and approved by the institutional review boards at the University of California, Davis. All data followed current laws, regulations, and IRB guidelines (such as sharing de-identified data that does not contain information used to establish the identity of individual deceased subjects). De-identified data does not contain personal health information (PHI) like names, social security numbers, addresses and phone numbers. Data were shared with a randomly generated pseudo-identification number.

Availability of data and material

The core datasets used and/or analyzed during the current study are openly available (see **Data Availability**). Further materials are available from the corresponding author on reasonable request.

Funding

This study was funded by a NIH P30 AG010129 grant (BND, CD, LWJ, and LB), a Paul G. Allen Family Foundation Distinguished Investigator Award (MJK), and the China Scholarship Council (ZT). These agencies had no role in any aspect of the study, including study design, data collection, analysis, or writing.

References

1. Dickson, D. W. The pathogenesis of senile plaques. *J. Neuropathol. Exp. Neurol.* **56,** 321–339 (1997).
2. Dugger, B. N. & Dickson, D. W. Pathology of Neurodegenerative Diseases. *Cold Spring Harb. Perspect. Biol.* **9,** (2017).
3. Hyman, B. T. *et al.* National Institute on Aging-Alzheimer's Association guidelines for the neuropathologic assessment of Alzheimer's disease. *Alzheimers. Dement.* **8,** 1–13 (2012).
4. Mirra, S. S. *et al.* The Consortium to Establish a Registry for Alzheimer's Disease (CERAD). Part II. Standardization of the neuropathologic assessment of Alzheimer's disease. *Neurology* **41,** 479–486 (1991).
5. Thal, D. R., Rub, U., Orantes, M. & Braak, H. Phases of A beta-deposition in the human brain and its relevance for the development of AD. *Neurology* **58,** 1791–1800 (2002).
6. Jack, C. R. *et al.* Hypothetical model of dynamic biomarkers of the Alzheimer's pathological cascade. *Lancet Neurol.* **9,** 119–128 (2010).
7. 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).
8. 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,** 1–11 (2012).
9. 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).
10. 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).
11. Greenberg, S. M. & Vonsattel, J. P. G. Diagnosis of cerebral amyloid angiopathy - Sensitivity and specificity of cortical biopsy. *Stroke* **28,** 1418–1422 (1997).
12. Vonsattel, J. P. *et al.* Cerebral amyloid angiopathy without and with cerebral hemorrhages: a comparative histological study. *Ann. Neurol.* **30,** 637–649 (1991).
13. Mirra, S. S. *et al.* Interlaboratory comparison of neuropathology assessments in Alzheimer's disease: a study of the Consortium to Establish a Registry for Alzheimer's Disease (CERAD). *J. Neuropathol. Exp. Neurol.* **53,** 303–315 (1994).
14. Iwatsubo, T. *et al.* Visualization of A beta 42(43) and A beta 40 in senile plaques with end-specific A beta monoclonals: evidence that an initially deposited species is A beta 42(43). *Neuron* **13,** 45–53 (1994).
15. Shinkai, Y. *et al.* Amyloid beta-protein deposition in the leptomeninges and cerebral cortex. *Ann. Neurol.* **42,** 899–908 (1997).
16. Dickson, T. C. & Vickers, J. C. The morphological phenotype of beta-amyloid plaques and associated

neuritic changes in Alzheimer's disease. *Neuroscience* **105,** 99–107 (2001).

17. Murray, M. E. *et al.* Neuropathologically defined subtypes of Alzheimer's disease with distinct clinical characteristics: a retrospective study. *Lancet Neurol.* **10,** 785–796 (2011).
18. Dugger, B. N. *et al.* Rapid eye movement sleep behavior disorder and subtypes in autopsy-confirmed dementia with Lewy bodies. *Mov. Disord.* **27,** 72–78 (2012).
19. Lewis, J. & Dickson, D. W. Propagation of tau pathology: hypotheses, discoveries, and yet unresolved questions from experimental and human brain studies. *Acta Neuropathol.* **131,** 27–48 (2016).
20. Neltner, J. H. *et al.* Digital pathology and image analysis for robust high-throughput quantitative assessment of Alzheimer disease neuropathologic changes. *J. Neuropathol. Exp. Neurol.* **71,** 1075–1085 (2012).
21. Taylor, C. R. & Levenson, R. M. Quantification of immunohistochemistry--issues concerning methods, utility and semiquantitative assessment II. *Histopathology* **49,** 411–424 (2006).
22. Howard, V. & Reed, M. *Unbiased stereology: three-dimensional measurement in microscopy*. (Garland Science, 2004).
23. Perl, D. P. *et al.* Practical approaches to stereology in the setting of aging- and disease-related brain banks. *J. Chem. Neuroanat.* **20,** 7–19 (2000).
24. Alafuzoff, I. *et al.* Staging of neurofibrillary pathology in Alzheimer's disease: a study of the BrainNet Europe Consortium. *Brain Pathol.* **18,** 484–496 (2008).
25. Shen, D., Wu, G. & Suk, H.-I. Deep Learning in Medical Image Analysis. *Annu. Rev. Biomed. Eng.* **19,** 221–248 (2017).
26. Litjens, G. *et al.* A survey on deep learning in medical image analysis. *Med. Image Anal.* **42,** 60–88 (2017).
27. Esteva, A. *et al.* Dermatologist-level classification of skin cancer with deep neural networks. *Nature* **542,** 115–118 (2017).
28. Khosravi, P., Kazemi, E., Imielinski, M., Elemento, O. & Hajirasouliha, I. Deep Convolutional Neural Networks Enable Discrimination of Heterogeneous Digital Pathology Images. *EBioMedicine* **27,** 317–328 (2018).
29. Ehteshami Bejnordi, B. *et al.* Diagnostic Assessment of Deep Learning Algorithms for Detection of Lymph Node Metastases in Women With Breast Cancer. *JAMA* **318,** 2199–2210 (2017).
30. Lu, D. *et al.* Multimodal and Multiscale Deep Neural Networks for the Early Diagnosis of Alzheimer's Disease using structural MR and FDG-PET images. *Sci. Rep.* **8,** 5697 (2018).
31. Beebe-Wang, N., Celik, S. & Lee, S.-I. MD-AD: Multi-task deep learning for Alzheimer's disease neuropathology. *bioRxiv* 331942 (2018). doi:10.1101/331942
32. Litjens, G. *et al.* Deep learning as a tool for increased accuracy and efficiency of histopathological diagnosis. *Sci. Rep.* **6,** 26286 (2016).
33. Lou, Y., Caruana, R. & Gehrke, J. Intelligible Models for Classification and Regression. in *Proceedings of*

the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining 150–158 (ACM, 2012).

34. Samek, W., Wiegand, T. & Müller, K.-R. Explainable Artificial Intelligence: Understanding, Visualizing and Interpreting Deep Learning Models. *arXiv [cs.AI]* (2017).
35. Krizhevsky, A., Sutskever, I. & Hinton, G. E. ImageNet Classification with Deep Convolutional Neural Networks. in *Advances in Neural Information Processing Systems 25* (eds. Pereira, F., Burges, C. J. C., Bottou, L. & Weinberger, K. Q.) 1097–1105 (Curran Associates, Inc., 2012).
36. Documentation Team. *AWS Elastic Beanstalk Developer Guide*. (Samurai Media Limited, 2018).
37. Ciregan, D., Meier, U. & Schmidhuber, J. Multi-column deep neural networks for image classification. in *2012 IEEE Conference on Computer Vision and Pattern Recognition* 3642–3649 (2012).
38. Simard, P. Y., Steinkraus, D. & Platt, J. C. Best practices for convolutional neural networks applied to visual document analysis. in *Seventh International Conference on Document Analysis and Recognition, 2003. Proceedings.* 958–963 (2003).
39. He, H. & Garcia, E. A. Learning from Imbalanced Data. *IEEE Trans. Knowl. Data Eng.* 1263–1284 (2008).
40. Wang, T., Wu, D. J., Coates, A. & Ng, A. Y. End-to-end text recognition with convolutional neural networks. in *Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012)* 3304–3308 (2012).
41. Vinters, H. V. Cerebral amyloid angiopathy. A critical review. *Stroke* **18,** 311–324 (1987).
42. Iwamoto, N., Nishiyama, E., Ohwada, J. & Arai, H. Distribution of amyloid deposits in the cerebral white matter of the Alzheimer's disease brain: relationship to blood vessels. *Acta Neuropathol.* **93,** 334–340 (1997).
43. Cruz-Roa, A. *et al.* Accurate and reproducible invasive breast cancer detection in whole-slide images: A Deep Learning approach for quantifying tumor extent. *Sci. Rep.* **7,** 46450 (2017).
44. Selvaraju, R. R. *et al.* Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. *arXiv [cs.CV]* (2016).
45. Zeiler, M. D. & Fergus, R. Visualizing and Understanding Convolutional Networks. in *Computer Vision – ECCV 2014* (eds. Fleet, D., Pajdla, T., Schiele, B. & Tuytelaars, T.) **8689,** 818–833 (Springer International Publishing, 2014).
46. Komura, D. & Ishikawa, S. Machine Learning Methods for Histopathological Image Analysis. *Comput. Struct. Biotechnol. J.* **16,** 34–42 (2018).
47. Ciompi, F. *et al.* The importance of stain normalization in colorectal tissue classification with convolutional networks. in *2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017)* 160–163 (2017).
48. Marée, R. *et al.* Collaborative analysis of multi-gigapixel imaging data using Cytomine. *Bioinformatics* **32,** 1395–1401 (2016).
49. Nalisnik, M. *et al.* Interactive phenotyping of large-scale histology imaging data with HistomicsML. *Sci.*

Rep. **7,** 14588 (2017).

50. Deterding, S., Sicart, M., Nacke, L., O'Hara, K. & Dixon, D. Gamification. Using Game-design Elements in Non-gaming Contexts. in *CHI '11 Extended Abstracts on Human Factors in Computing Systems* 2425–2428 (ACM, 2011).
51. Deterding, S., Dixon, D., Khaled, R. & Nacke, L. From game design elements to gamefulness: defining gamification. in *Proceedings of the 15th international academic MindTrek conference: Envisioning future media environments* 9–15 (ACM, 2011).
52. Seaborn, K. & Fels, D. I. Gamification in theory and action: A survey. *Int. J. Hum. Comput. Stud.* **74,** 14–31 (2015).
53. Hamari, J. & Eranti, V. Framework for Designing and Evaluating Game Achievements. *Digra conference* (2011).
54. Simonyan, K. & Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv [cs.CV]* (2014).
55. Cullen, K. M., Kócsi, Z. & Stone, J. Microvascular pathology in the aging human brain: evidence that senile plaques are sites of microhaemorrhages. *Neurobiol. Aging* **27,** 1786–1796 (2006).
56. Majocha, R. E., Benes, F. M., Reifel, J. L., Rodenrys, A. M. & Marotta, C. A. Laminar-specific distribution and infrastructural detail of amyloid in the Alzheimer disease cortex visualized by computer-enhanced imaging of epitopes recognized by monoclonal antibodies. *Proc. Natl. Acad. Sci. U. S. A.* **85,** 6182–6186 (1988).
57. Chuang, K. V. & Keiser, M. J. Adversarial Controls for Scientific Machine Learning. *ACS Chem. Biol.* **13,** 2819–2821 (2018).
58. Dugger, B. N. *et al.* Presence of Striatal Amyloid Plaques in Parkinson's Disease Dementia Predicts Concomitant Alzheimer's Disease: Usefulness for Amyloid Imaging. *J. Parkinsons. Dis.* **2,** 57–65 (2012).
59. Serrano-Pozo, A., Frosch, M. P., Masliah, E. & Hyman, B. T. Neuropathological alterations in Alzheimer disease. *Cold Spring Harb. Perspect. Med.* **1,** a006189 (2011).
60. Vinters, H. V. *et al.* Review: Vascular dementia: clinicopathologic and genetic considerations. *Neuropathol. Appl. Neurobiol.* **44,** 247–266 (2018).
61. Sharma, G. & Carter, A. Artificial Intelligence and the Pathologist: Future Frenemies? *Arch. Pathol. Lab. Med.* **141,** 622–623 (2017).
62. pyvips. Available at: https://github.com/libvips/pyvips. (Accessed: 5th October 2018)
63. Reinhard, E., Ashikhmin, N., Gooch, B. & Shirley, P. Color transfer between images. *IEEE Comput. Graph. Appl.* **21,** 34–41 (2001).
64. Bradski, G. & Others. The opencv library (2000). *Dr. Dobb's Journal of Software Tools* (2000).
65. Paszke, A., Gross, S., Chintala, S. & Chanan, G. PyTorch. (2017). Available at: http://pytorch.org.
66. Kingma, D. P. & Ba, J. Adam: A Method for Stochastic Optimization. *arXiv [cs.LG]* (2014).
67. Ozbulak, U. pytorch-cnn-visualizations. Available at:

https://github.com/utkuozbulak/pytorch-cnn-visualizations. (Accessed: 5th October 2018)

68. Jones E, Oliphant E, Peterson P, et al. SciPy: Open Source Scientific Tools for Python. (2001). Available at: http://www.scipy.org/. (Accessed: 18th October 2018)

Acknowledgements

The authors thank the families and participants of the UCD-ADC for their generous donations. The authors would also like to thank the Department of Pathology and Laboratory Medicine at the University of California Davis for the use of their Aperio Scan scope and Kelsey Erickson and Justin Athey for aiding with slide scanning.

Author contributions

The studies were conceptualized, results analyzed, and manuscript drafted by ZT, KVC, MJK, and BND. ZT and KVC wrote the code, performed the computational experiments, trained the models, and wrote the first draft of the manuscript, with guidance from MJK and BND. CD, LWJ, and BND provided clinical and neuropathology work-up for case classifications and data interpretation. LB provided supervision of statistical analysis and interpretation of data. BND provided case selection and annotations, and neuropathology data interpretation. BND and MJK provided overall study supervision. All authors were involved in critical revisions of the manuscript, and have read and approved the final version.

Competing interests

The authors declare no competing interests related to this study. Dr. DeCarli is a consultant to Novartis. Dr. Keiser is a consultant to Daiichi Sankyo. Dr. Dugger has received previous funding from Daiichi Sankyo unrelated to this project.

Table and Figure Captions

Table 1. Summary of annotated plaque tile dataset by project phase. Remaining unlisted percentages correspond to “not sure” or “flagged.”

Table 2. Annotated plaque dataset distribution by class and model performance.

Figure 1. a. Current protocols for neuropathological assessment of WSIs typically rely on comparatively coarse-grained semi-quantitative scoring such as CERAD[4]. **b**. We report an automated computational approach to process entire digitized immunohistochemical stained archival slides, leveraging convolutional neural networks for amyloid plaque classification and localization.

Figure 2. Creation of an annotated plaque and CAA dataset for machine learning. **a.** Summary of image processing pipeline, including color normalization, immunohistochemical (IHC) stain segmentation, and extraction of candidate objects (red boxes), followed by rapid expert annotation using a cloud-based web application. **b.** Examples of extracted cored plaques (top row), diffuse plaques (middle), and CAA (bottom) and their surrounding tissue area. Rectangles shown in red bound the candidate object during the labeling process. Scale bar=25 um. **c.** A custom web interface allows for the rapid annotation of plaques by mouse or keystroke, with visualization of raw (without color adjustments) and normalized images, showing the object bounding boxes around which the tile is automatically centered and cropped.

Figure 3. CNN models identify three amyloid plaque types in image tiles. **a.** The optimized CNN model architecture contained six convolutional layers and two dense layers, using exclusively 3x3 kernels and alternating max-pooling layers. **b.** Examples of correct CNN predictions. The ground truth “expert label” row indicates the pathologies that had been manually found within the tile image. The “predicted” row shows corresponding model confidences for cored plaque (yellow arrow), diffuse plaque (red), and CAA (blue) classes (from left to right). Model predictions range from 0.00 to 1.00, where a higher score indicates a higher predicted confidence by the CNN for that plaque class (e.g., the 1.00 corresponds to 100% model confidence that a cored plaque is present in the leftmost panel). **c.** Examples of CNN predictions that do not agree with the expert manual annotation. Incorrect model predictions are indicated by light orange backgrounds in the “predicted” column; green backgrounds correspond to correct predictions. Scale bar=25 um for all images.

Figure 4. Predictive performance on the held-out Phase-II test set. **a**. Receiver operator characteristic (ROC) and **b.** Precision-recall curves (PRC) for cored (magenta lines) and diffuse (blue lines) plaques. Blue star marks the best trade-off point where prediction confidence threshold equals 0.93. **c.** Summarized areas under the ROC and PRC (AUPRC and AUROC) of five independently-trained CNNs per point for the task of cored

plaque classification, as a function of training dataset size. Dataset was randomly subsetted at each point independent of the date of tile annotation. **d.** AUPRC and AUROC of CNNs for the same task of cored-plaque classification, as a function of chronological dataset growth by annotation timestamp, over the course of the project, showing chronology-dependent dataset effects.

Figure 5. Prediction confidence heatmaps for cored plaques, diffuse plaques, and CAAs. **a**. Whole slide overview visualization, revealing broad amyloid distribution patterns. Scale bar=3 mm. **b.** Higher magnification (4x) view of blue-boxed region from panel (a). Scale bar=750 um. **c.** Higher magnification (20x) view of blue-boxed region from panel (b). Green box marks cored plaque manual annotation. Scale bar=150 um. Confidence scales for each panel were bottom-capped to aid visualization, such that only confidence scores ≥0.8 are plotted. Approximate (hand-drawn) boundaries of gray versus white matter (dotted line) and tissue boundaries (solid lines) are overlaid for reference.

Figure 6. Visualization of a representative example from the cored-plaque classification tests plotted in **Fig. 4**. **a**. CNN model prediction confidence maps (middle panel, as in **Fig. 5c**) overlaid onto the original slide tile. Bounding boxes mark cored-plaque expert annotations (left panel, green box). The combined map (right panel) assesses agreement between the model's predictions and the expert labels, where pixels are colored by a semi-transparent overlay as true positive (blue), false positive (orange), true negative (cyan), and false negative (red) areas. Scale bar=150 um. **b**. Prediction-versus-annotation agreement map generated as in (a), but with a larger field for greater tissue and plaque clustering context. Scale bar=750 um.

Figure 7. Model interpretability studies using machine-learning introspection techniques. **a**. A cored plaque example (top row, yellow arrow). For the task of cored-plaque prediction, the activation map (by guided grad-CAM; left, second row) and the feature occlusion map (right, second row) identify the amyloid core (yellow arrow) as the defining morphological feature. By contrast, the diffuse stained region (red arrow) only arises as a salient feature during diffuse-plaque and CAA prediction tasks (third and fourth rows, respectively). **b.** Diffuse plaque example where activation and feature occlusion maps focus on ill-defined amorphous amyloid contours for diffuse-plaque classification task (third row). **c.** CAA example, where the CAA task's activation and feature occlusion maps (fourth row) highlight amyloid ring pixels within the media of the cortical vessel (blue arrow), while for cored and diffuse tasks the small punctate IHC staining is considered salient (red arrow; second and third rows). **d.** Example containing both diffuse (red arrow) and cored (yellow arrow) plaques in same tile illustrate the difference between activation and feature occlusion maps. Scale bar=25 um.

Figure 8. Comparison of CNN-based plaque-burden scores versus manual CERAD-like semi-quantitative scores at a whole-slide level. **a.** The automatic and manual scores correlate well across the entire dataset of 62 WSIs, comprising the original Phase I-III slide set plus 20 additional blinded WSIs. **b.** Correlations assessed

only on the 20 blinded WSIs not used in any previous step of the study. Box plots show median (band inside the box), interquartile range (IQR, box), minimums and maximums within 1.5 times the IQR (whiskers), and outliers (points beyond the whiskers), with a dot per WSI. Matrices in the second row of each panel exhaustively plot *p*-values of CNN-based score distributions between all pairs of CERAD-like categories for the corresponding box plot, where squares are colored in log scale by *p*<1e-4 (blue) to *p*=0.05 (gray) to *p*=1 (red; insignificant) using Student's *t*-tests.

Tables

Table 1.

Phase	Cored plaque	Diffuse plaque	CAA	Total
Development Phase I	1,233 (2.24 %)	46,650 (84.82 %)	778 (1.14 %)	55,001
Development Phase II	1,035 (9.38 %)	7,610 (69.00 %)	1,405 (12.74 %)	11,029
Development Total	2,268 (3.43 %)	54,260 (82.17 %)	2,183 (3.31 %)	66,030
Test (Phase III)	83 (0.76 %)	10,234 (94.12 %)	7 (0.06 %)	10,873

Table 2.

	Images	Cored plaque	Diffuse plaque	CAA	AUROC (cored)	AUPRC (cored)
Train	61,370	2,141	48,123	2227	0.993	0.981
Validation	8,630	381	7,487	126	0.983	0.842
Test	10,873	98	10,480	7	0.993	0.744

Figure 1

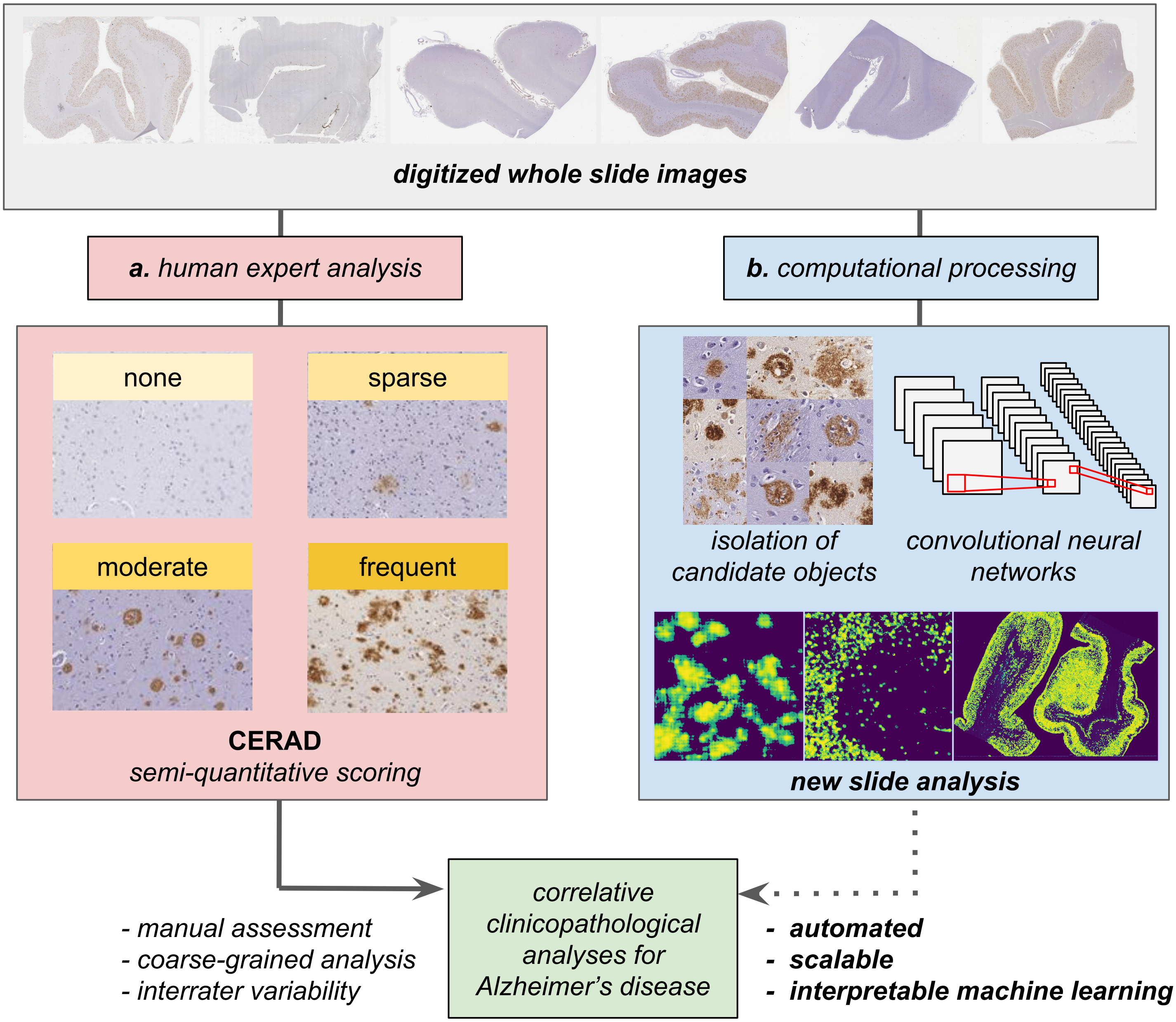

digitized whole slide images
a. human expert analysis
b. computational processing
none
sparse
moderate
frequent
CERAD
semi-quantitative scoring
isolation of candidate objects
convolutional neural networks
new slide analysis
correlative clinicopathological analyses for Alzheimer's disease
- manual assessment
- coarse-grained analysis
- interrater variability
- automated
- scalable
- interpretable machine learning

Figure 2

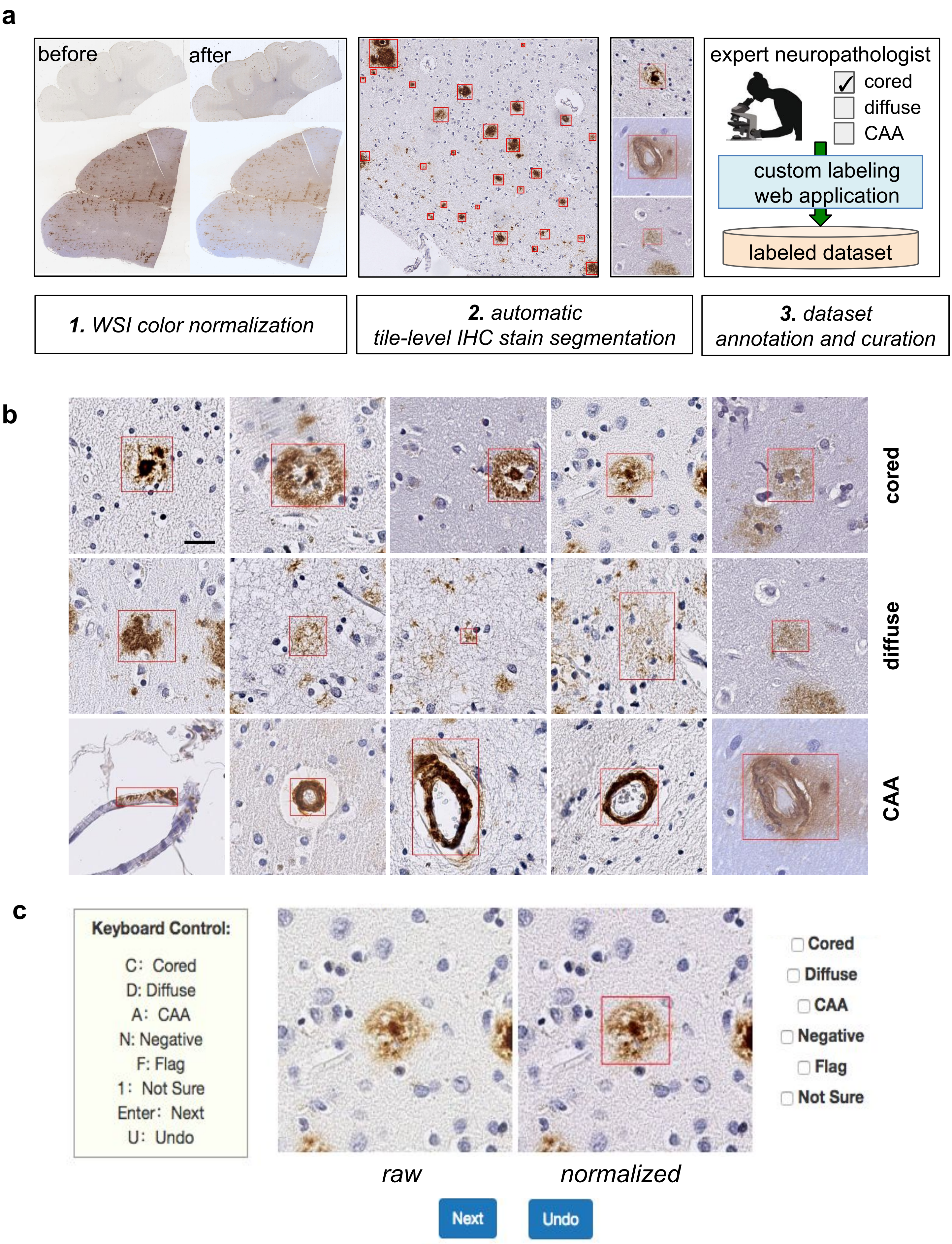

a
before
after
expert neuropathologist
☑ cored
□ diffuse
□ CAA
custom labeling web application
labeled dataset
1. WSI color normalization
2. automatic tile-level IHC stain segmentation
3. dataset annotation and curation
b
cored
diffuse
CAA
c
Keyboard Control:
C: Cored
D: Diffuse
A: CAA
N: Negative
F: Flag
1: Not Sure
Enter: Next
U: Undo
raw
normalized
□ Cored
□ Diffuse
□ CAA
□ Negative
□ Flag
□ Not Sure
Next
Undo

Figure 3

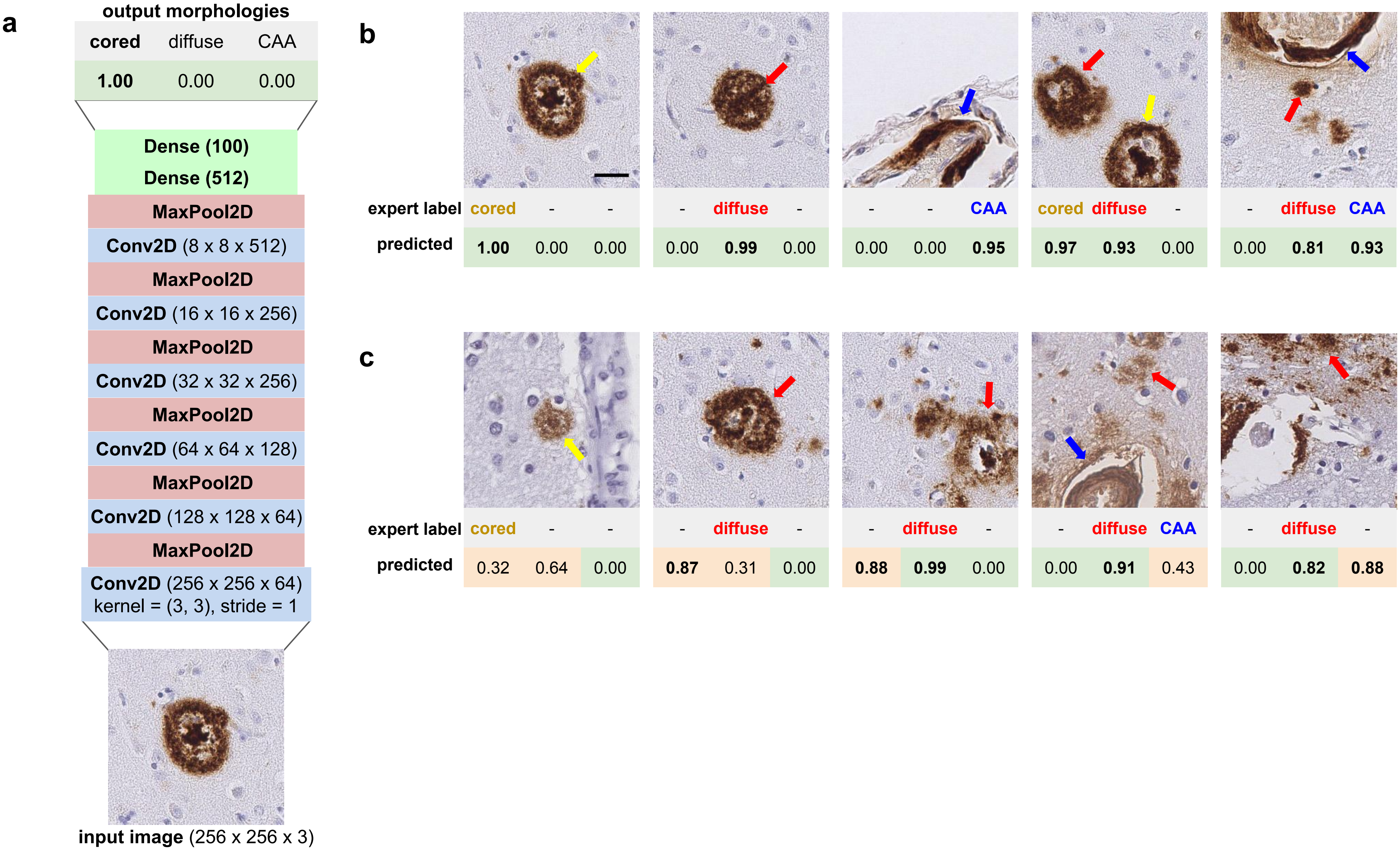

a
output morphologies
cored diffuse CAA
1.00 0.00 0.00
Dense (100)
Dense (512)
MaxPool2D
Conv2D (8 x 8 x 512)
MaxPool2D
Conv2D (16 x 16 x 256)
MaxPool2D
Conv2D (32 x 32 x 256)
MaxPool2D
Conv2D (64 x 64 x 128)
MaxPool2D
Conv2D (128 x 128 x 64)
MaxPool2D
Conv2D (256 x 256 x 64)
kernel = (3, 3), stride = 1
input image (256 x 256 x 3)
b
expert label cored - - | - diffuse - | - - CAA | cored diffuse - | - diffuse CAA
predicted 1.00 0.00 0.00 | 0.00 0.99 0.00 | 0.00 0.00 0.95 | 0.97 0.93 0.00 | 0.00 0.81 0.93
c
expert label cored - - | - diffuse - | - diffuse - | - diffuse CAA | - diffuse -
predicted 0.32 0.64 0.00 | 0.87 0.31 0.00 | 0.88 0.99 0.00 | 0.00 0.91 0.43 | 0.00 0.82 0.88

Figure 4

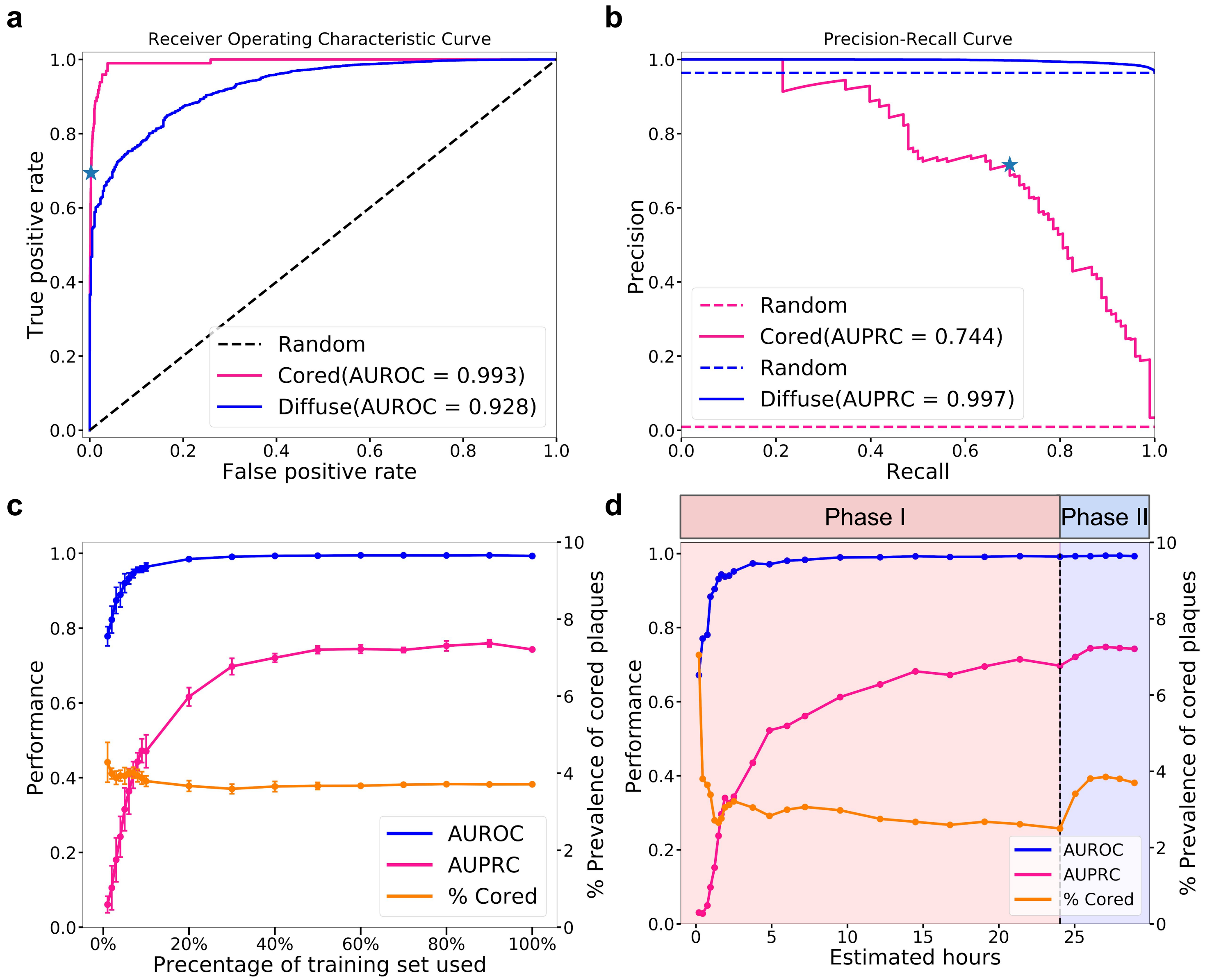

a
Receiver Operating Characteristic Curve
True positive rate
False positive rate
0.0
0.2
0.4
0.6
0.8
1.0
Random
Cored(AUROC = 0.993)
Diffuse(AUROC = 0.928)
b
Precision-Recall Curve
Precision
Recall
Random
Cored(AUPRC = 0.744)
Random
Diffuse(AUPRC = 0.997)
c
Performance
% Prevalence of cored plaques
Precentage of training set used
0%
20%
40%
60%
80%
100%
0
2
4
6
8
10
AUROC
AUPRC
% Cored
d
Phase I
Phase II
Performance
% Prevalence of cored plaques
Estimated hours
0
5
10
15
20
25
AUROC
AUPRC
% Cored

Figure 5

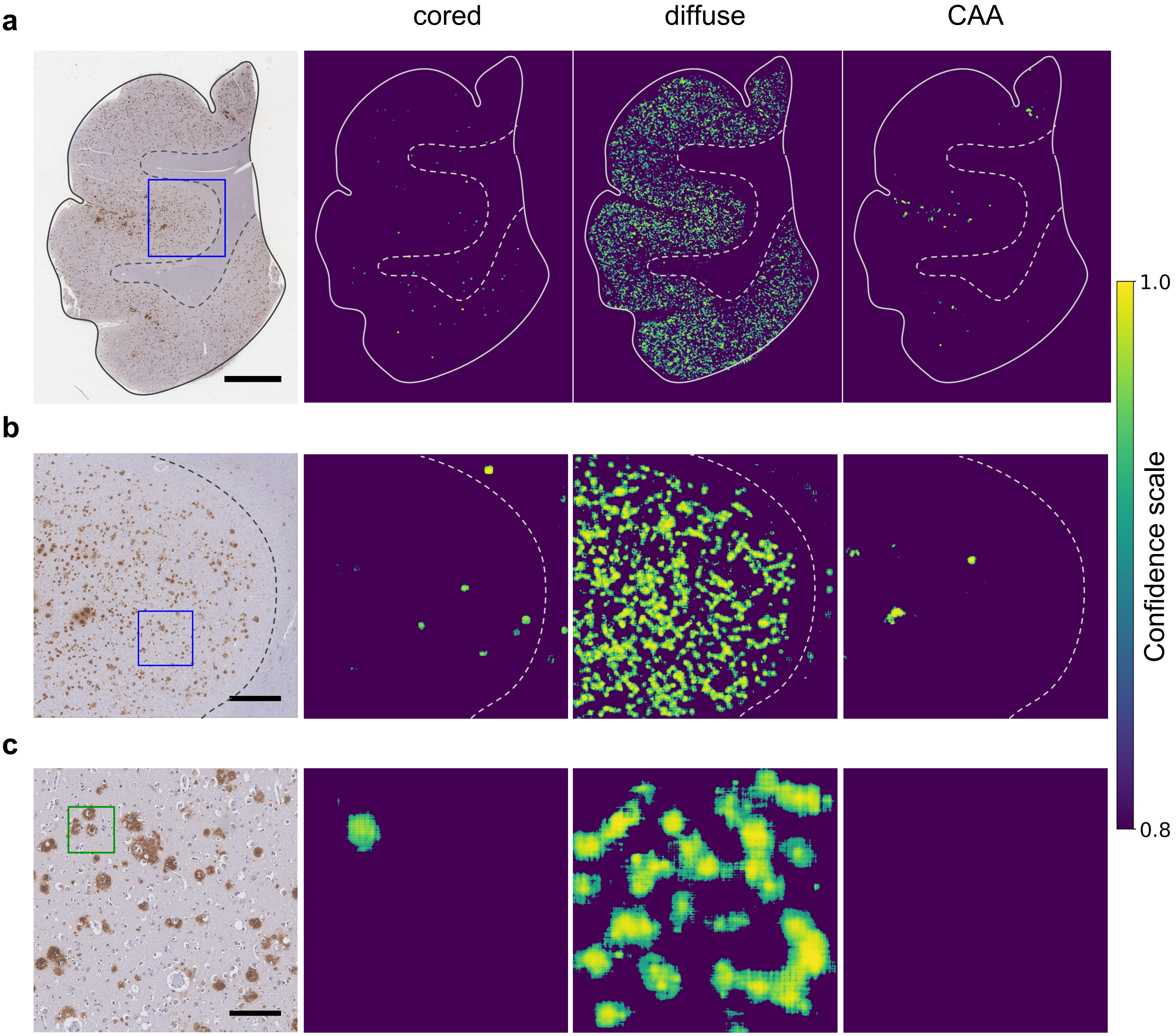


Figure 6

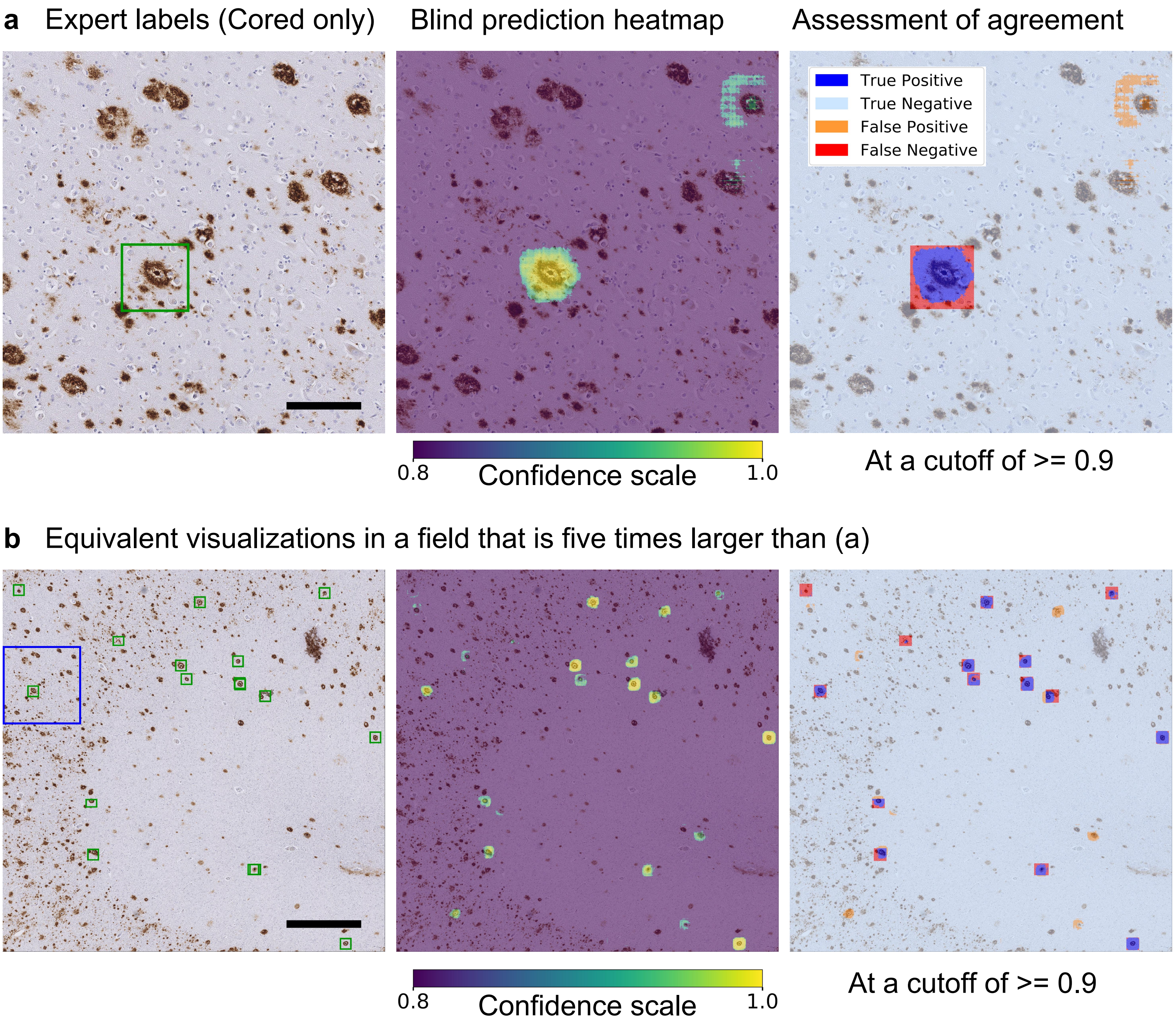
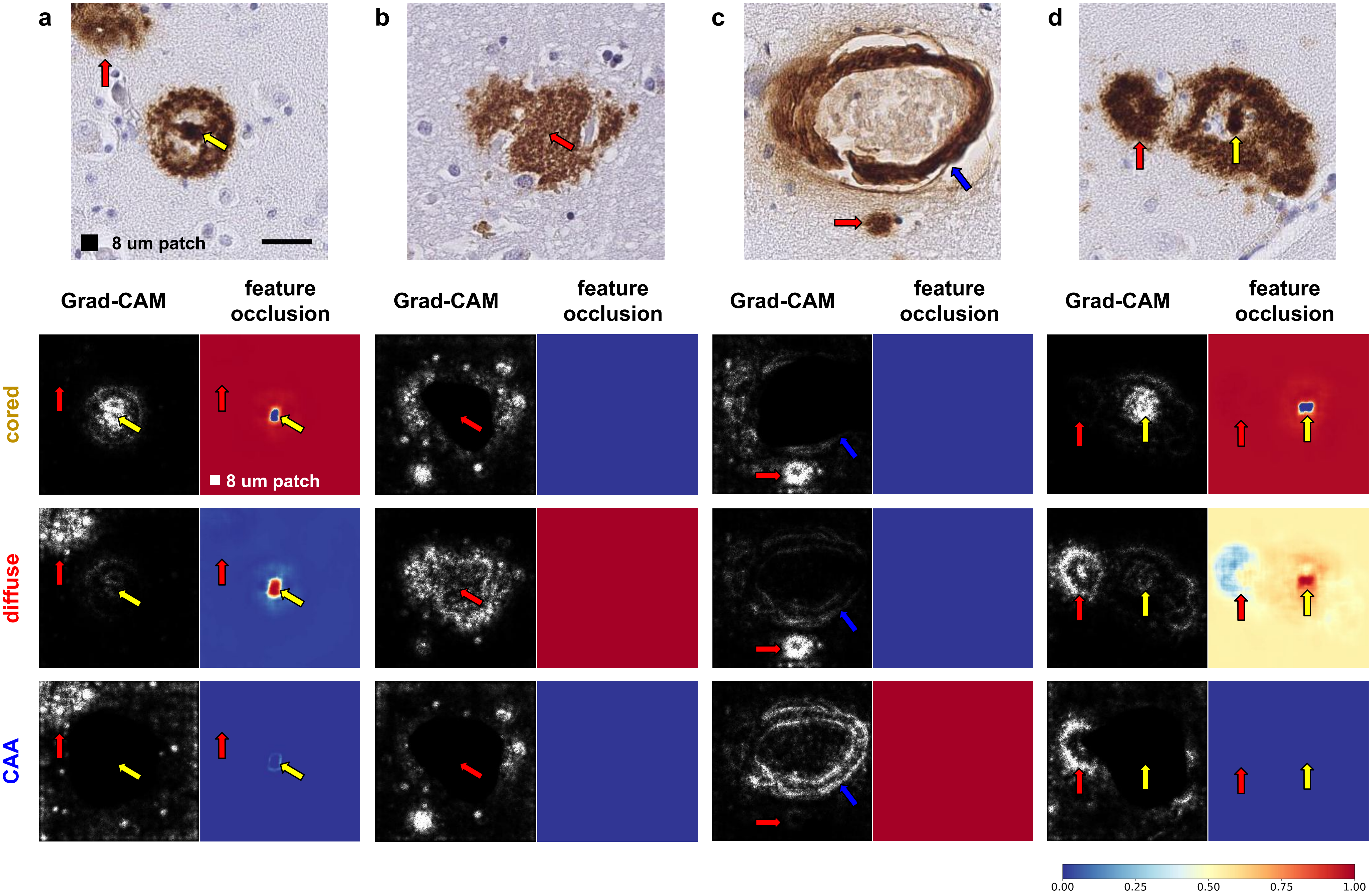
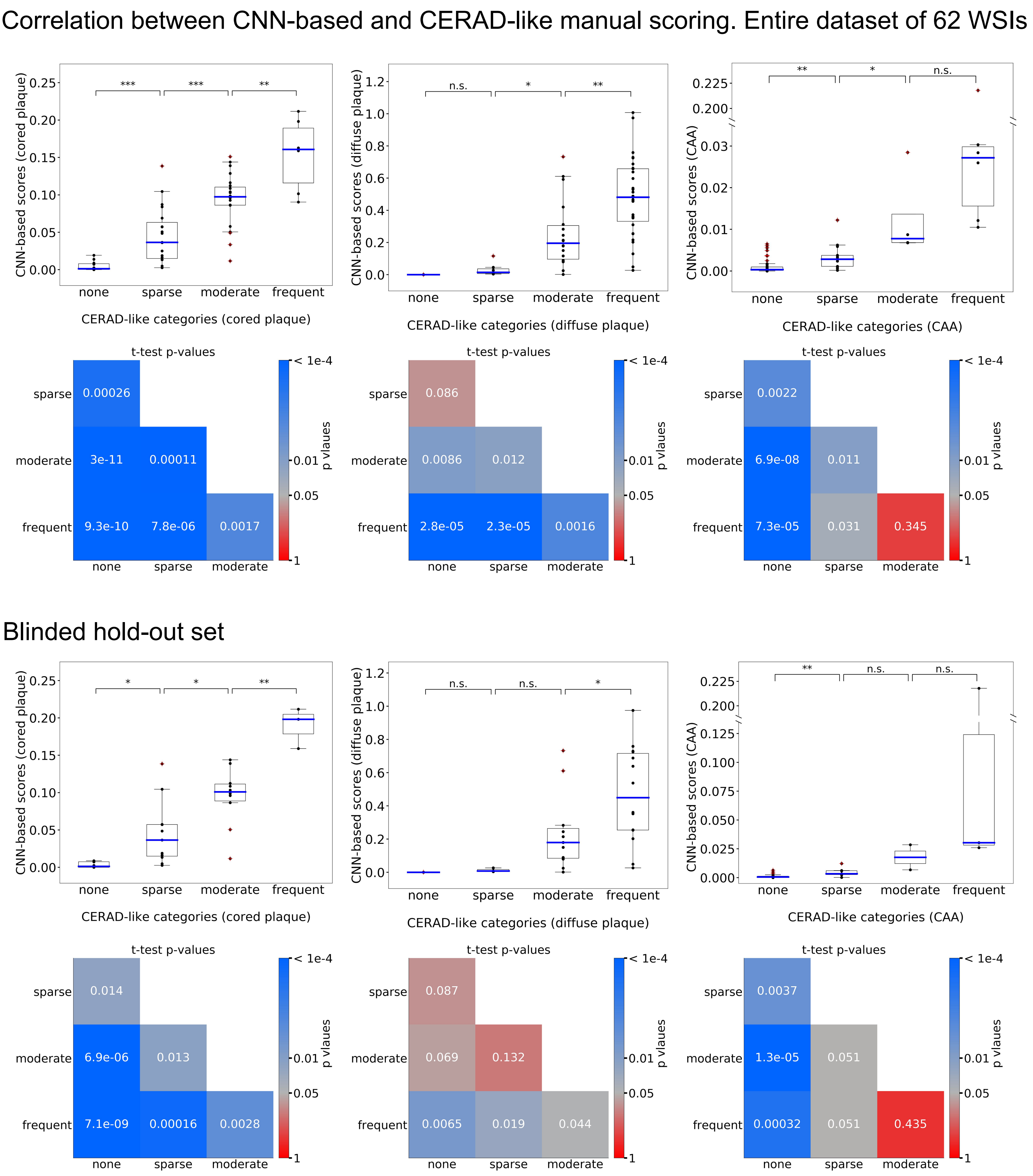

a
Expert labels (Cored only)
Blind prediction heatmap
Assessment of agreement
True Positive
True Negative
False Positive
False Negative
0.8
Confidence scale
1.0
At a cutoff of >= 0.9
b
Equivalent visualizations in a field that is five times larger than (a)
0.8
Confidence scale
1.0
At a cutoff of >= 0.9

Figure 7



a
8 um patch
b
c
d
Grad-CAM
feature occlusion
Grad-CAM
feature occlusion
Grad-CAM
feature occlusion
Grad-CAM
feature occlusion
cored
8 um patch
diffuse
CAA
0.00
0.25
0.50
0.75
1.00

Figure 8



a. Correlation between CNN-based and CERAD-like manual scoring. Entire dataset of 62 WSIs

b. Blinded hold-out set