Title

Preanalytic variable effects on segmentation and quantification machine learning algorithms for amyloid-β analyses on digitized human brain slides

Journal

Journal of Neuropathology & Experimental Neurology, 82(3)

ISSN

0022-3069

Authors

Oliveira, Luca Cerny
Lai, Zhengfeng
Harvey, Danielle
et al.

Publication Date

2023-02-21

DOI

10.1093/jnen/nlac132

Pre-analytic variable effects on segmentation and quantification machine learning algorithms for amyloid beta analyzes on digitized human brain slides

Luca Cerny Oliveira BS (1), Zhengfeng Lai MS (1), Danielle Harvey PhD (2), Kevin Nzenkue BS (3), Lee-Way Jin MD PhD (3), Charles DeCarli MD (4), Chen-Nee Chuah PhD (1), and Brittany N. Dugger PhD (3) *

1. Department of Electrical and Computer Engineering, University of California Davis, Davis, CA 95616, USA
2. Department of Public Health Sciences, University of California, Davis, Davis, CA 95616, USA
3. Department of Pathology and Laboratory Medicine, University of California Davis, Sacramento, CA 95817, USA
4. Department of Neurology, University of California, Davis, Sacramento, CA 95817, USA

* Corresponding author

Corresponding Author Address

Brittany N. Dugger, PhD
University of California Davis, Department of Pathology and Laboratory Medicine
4645 2nd Ave.
3400A Research Building III
Sacramento, CA 95817
phone: (916)734-3855
fax: (916) 734-2698
email: bndugger@ucdavis.edu

Running title: Pre-analytic effects on Digital Pathology

Keywords: Machine Learning, Deep learning, Neuropathology, Digital Pathology, Magnification, Slide scanner; whole slide image; Alzheimer's Disease

Sources of Support

The authors thank the families and participants of the University of California Davis Alzheimer's Disease Research Centers (ADRC) for their generous donations as well as ADRC staff and faculty for their contributions. Resources for this study were funded in part by grants from the National Institute on Aging of the National Institutes of Health under Award Numbers R01AG062517, P30AG072972, and R01AG056519, and a research grant from the California Department Of Public Health (19-10611) with partial funding from the 2019 California Budget Act. The views and opinions expressed in this manuscript are those of the author and do not necessarily reflect the official policy or position of any public health agency of California or of the United States government. We also thank the UC Davis Health Department of Pathology and Laboratory medicine as well as the laboratory of Dr. Alexander "Sandy" Borowsky for the use of digital slide scanners.

Conflicts of Interest Authors have no conflicts of interest to disclose related to this work.

Abstract
Computational machine learning-based frameworks could be advantageous for the field of neuropathology for scalable analyses. A recent deep learning (DL) framework has shown promise in automating the process of visualizing and quantifying different types of Amyloid-β deposits as well as segmenting white matter (WM) from gray matter (GM) on digitized immunohistochemically stained slides. However, this framework has only been trained and evaluated on Amyloid-β-stained slides with minimal changes in pre-analytic variables. In the current study, we evaluated select pre-analytical variables including magnification, compression rate, and storage format using three digital slides scanners (Zeiss Axioscan Z1, Leica Aperio AT2, and Leica Aperio GT 450) on over 60 whole slide images, in a cohort of 14 cases having a spectrum of Amyloid-β depositions. We conducted statistical comparisons of pre-analytic variables with Repeated Measures Analysis of Variance evaluating the outputs of two DL frameworks for segmentation and object classification tasks. For both WM/GM segmentation and Amyloid-β plaque classification tasks, there were statistical differences with respect to scanner types ($Ps < 0.05$) and magnifications ($Ps < 0.05$). This pilot study, although presenting small numbers of cases, highlights the significance of pre-analytic variables that may alter performance of machine learning algorithms.

Introduction

Whole-slide imaging has become an increasingly popular modality to assess brain tissues. With the help of a digital slide scanner, ultra-high resolution whole slide images (WSIs) are generated to aid in preservation of tissue details (1). WSIs can be viewed and annotated through computer software (such as Aperio ImageScope, ZEN, and QuPath)(2). The digitization of tissue information allows for application of computational approaches, which include but are not limited to machine learning (ML) and image processing that can aid with automated analyses of tissue.

Many types of pathologies exist within the brain, of which define the neuropathologic classification of many neurodegenerative diseases (3). For example, Amyloid-β deposits, in the form of plaques, are a hallmark pathological feature of Alzheimer's Disease (AD) (3). It is becoming advantageous to have more quantitative assessments of these pathologies for deeper phenotyping that is paving the way for precision medicine approaches for these devastating diseases (4-6). The quantification of pathologies, such as Amyloid-β plaques, can be a time-consuming task that has been automated through Convolutional Neural Networks (CNN) (7), a type of ML framework. Other Deep Learning (DL) studies in pathology applied similar techniques for WSI analysis (7-9). However, the performance of the aforementioned CNN models has not been fully demonstrated for WSIs scanned under variable conditions (i.e. magnifications, compression rates, etc.).

Many promising DL-based studies in neuropathology utilize WSIs from a single scanner, with single WSI formatting settings (10-13). Such design choices lead to a study with little or no variation in pre-analytical variables such as image format, image compression rate, and scanner types. Despite displaying competitive performance, studies may not adequately assess generalizability; this lack of pre-analytical variable diversity could be a concern for the reported performance metrics (10-13). Research on organ tissues other than brain, has revealed a model trained with data from diverse settings, such as different scanners, outperform models trained with single-source WSIs (14). Although two WSIs from different scanners may look identical to a human evaluator, they may look distinct to the DL model due to the scanner, formatting applied to the scan, and/or the digital watermark left by preprocessing software. Different formatting settings or scanners introduce variables including but are not limited to compression standard, compression rate, storage format, and magnification. The pixel values such as µm/pixel, in each WSI may differ due to compression or other variables and yet display an identical image to the evaluator. Since DL models learn through backpropagation (15) and thus "see" the pixel values, not the overall picture like an expert, they may be affected by the change in these pre-analytic variables.

Concerning the performance of DL frameworks when presented data with different pre-analytical variables, few studies have tested and observed degradation in performance with different tissue areas and quantities (16-17), different scanners (18), and different class distributions (19). Another study has experimented with a single framework facing changes in storage format and architecture used (20). In real-world model deployment,

data from different scanners or generated with different scanning formatting will likely be evaluated by the model, leading to a concern about the performance metrics displayed in studies that demonstrate no variance in pre-analytical variables. Studies that present the same pre-analytical variables in its training and testing data do not adequately test its model's generalizability despite good reported performance metrics. Therefore, there may be many published DL models with good reported performance that can only replicate its good performance when fed data similar to their training set.

Our study seeks to provide a proof of concept in the neuropathology realm examining DL models' potential prediction effects of different pre-analytic variables. Our study aims to test the generalizability capacity of two DL models, one for segmentation the other for classification, trained using WSIs from a single scanner by testing them on the same slides scanned with different scanners and scanner settings. By displaying and comparing the qualitative and quantitative outputs of two different DL tasks when applied to data with different pre-analytical variables, we aim to report and highlight effects by such variable changes.

Methods

I. Datasets

We utilized WSI from a total of 14 cases from slides of formalin-fixed paraffin-embedded 5µm sections of post-mortem human brain temporal cortex immunohistochemically stained with an antibody against Amyloid-β diluted 1:1600 (4G8, BioLegend ,formally Covance, San Diego, CA), all sections were subjected to standard procedures on automated machines, pretreatment included 10mins in 87% formic acid, and endogenous peroxidases were block with 3% Hydrogen Peroxide. All antibody staining was conducted on an autostainer (DAKO AutostainerLink48, Agilent, Santa Clara, CA, USA) utilizing proper positive and negative control for each specific antibody. All staining was conducted using proper controls by the University of California Davis Histology Core, which is a Clinical Laboratory Improvement Amendments (CLIA) and College of American Pathologists (CAP) accredited laboratory that also operates under the best laboratory practices standards and meets all Federal, State of California and UC Davis guidelines and regulations. These stained slides were digitized to create six different WSIs datasets having different pre-analytic variables (see Figure 1 for schematic). In the following sections, we mention each of our datasets according to its pre-analytical variables. The names of the datasets reflect a formatting of Scanner-Magnification-Compression. To ensure fair performance through similar processing times for each WSI, all 40X WSIs were resized to 20X through PyVips package. Other than resizing, no pre-processing was done in any of the WSIs. No pre-processing was applied prior to feeding the WSIs to the model to avoid any digital watermarks generated by the image software such as ImageScope or ZEN. All scanners undergo routine servicing work once a year. The AT2 scanner was purchased in 2016, the Zeiss Axio Z1 Scanner was purchased in 2019, and the GT450 was purchased in 2021. Our work selected the standard processing method for all scanners, which include but is not limited to standard automatic color profile and

tissue detection. Tissue detection automatically crops the WSI to ensure reduced background. Color profile normalizes the pixel values for optimal monitor display. These effects of cropping and color profile can be observed in Supplemental Figure 1.

A. AT2-20X

A total of 14 slides from our cohort were scanned into JPEG-2000 compressed .svs files. The WSIs were digitized by Leica Aperio AT2 at 20X magnification. This dataset contains the same pre-analytical variables as the WSIs used for training of both Amyloid-β deposit detection and WM/GM segmentation.

B. AT2-40X

All 14 slides from our cohort scanned into JPEG-2000 compressed .svs files. The WSIs were digitized by Leica Aperio AT2 at 40X magnification. This dataset presents only one pre-analytical variable change, magnitude change, as the WSIs used for training of both evaluated DL frameworks. Case 11 displayed cover slip deahderance not seen in other datasets.

C. GT450-40X

A total of 14 slides from our cohort were scanned into JPEG-2000 compressed .svs files. The WSIs were digitized by Leica Aperio GT450 at 40X magnification. Despite having the same storage format, and compression standard, these WSIs came from a different Leica scan than the AT2.

D. Axio-Z1-40X-45

A total of 13 slides from our cohort were scanned into JPEG-XR compressed .czi files. The WSIs were digitized by Zeiss Axio Z1 scanner at 40X magnification. The JPEG-XR compression reduced the size of the file by 55%.

E. Axio-Z1-40X-75

A total of 13 slides from our cohort were scanned into JPEG-XR compressed .czi files. The WSIs were digitized by Zeiss Axio Z1 scanner at 40X magnification. The JPEG-XR compression reduced the size of the file by 25%.

II. Evaluated Pipelines

This study evaluated two pre-trained models (overall workflow is depicted in Figure 2). The first model, aimed at WM/GM segmentation (8), was trained on 20X JPEG-2000 compressed svs slides digitized from Leica Aperio AT2. The second model, aimed at detecting Amyloid-β plaques (7), was trained on 20X JPEG-2000 compressed svs slides. We performed no tuning or additional training on any of the two models. The pre-analytic variables for the data used in the two pre-trained models displayed constant scanner (Aperio AT2), constant magnification (20X), consistent storage format (SVS), and

constant compression standard (JPEG-2000), matching the pre-analytic variables from the AT2-20X dataset.

Both models employed CNN-based DL. The Amyloid-β deposit detection was originally trained on a version of VGG (21). The WM/GM segmentation model was trained on a version of ResNet-18 (22). Both ResNet and VGG are commonly used CNN-based DL architectures. The pipeline used to generate both models' predictions were similar to the one described in (23). We patched each WSI in 256x256 segments and those patches were the input to both classification and segmentation models simultaneously as pictured in Figure 2. Although the input is the same, each model performs different tasks, while the ResNet performs patch-based segmentation, the VGG model performs classification and detection of Amyloid-β present in each patch.

The ResNet-based WM/GM segmentation module outputs a heatmap with yellow, cyan, and black representing WM, GM and background respectively (as seen in Supplemental Figure 2). The model also outputs WM and GM size in um/pixel, which we use to calculate the WM/GM ratio. The VGG-based Amyloid-β deposit detection module outputs separate heatmaps based on each plaque classification (Cored and Diffuse) colored in red. Counts/area of each plaque classification were also generated by incorporating the WM/GM predictions. All code related to these processes are located in this GitHub (https://github.com/ucdrubinet/BrainSec).

III. Registration

Due to the distinct field-of-view (FOV) and automatic tissue detection present in each scanner, the output WSI files from different scanners (see supplemental Table 1 for additional details on each WSIs parameters) are not aligned and present different tissue sizes and aspect ratios despite being generated from the same slide. Automatic cropping caused loss of tissue area for select Zeiss scans (see supplemental figure 1 for example).

Hence, to register the WSIs and ensure as much alignment and as little loss of tissue as possible, we employed a technique for re-stained histological whole slide image co-registration (25). Aligning re-stained WSIs is similar to our task since the tissue borders are similar between re-stained slides. However, due to the difference in magnification in some WSIs, we also need to resize the 40X files into 20X to ensure similar tissue size. We achieved this by calculating the resizing factor that allowed for the height and width difference to be minimal when compared to the original AT2-20X WSI. Due to distinct aspect ratios from different scanners, the height and width from resized WSIs was not able to match the ones from the original AT2-20X WSI. This required an additional manual tuning step to the registration technique employed. All code related to these processes are located in this GitHub (https://github.com/smujiang/Re-stained_WSIs_Registration).

IV. Statistical Analysis

Because all slides were scanned using each of the scanners, repeated measures analysis of variance (ANOVA) was used to compare differences in WM/GM segmentation and Amyloid-β core or diffuse plaque counts derived from the ML models across pre-analytic variables. Key factors of interest included scanner, magnification, and compression rate. Not all combinations of factors were considered, so separate analyses were conducted for each comparison of interest, including all relevant data. For example, when considering compression rate, only outcomes from the slides on the Zeiss Axioscan were included. All analyses were conducted using Python and a p-value less than 0.05 was considered statistically significant.

Results

I. Effects of Pre-Analytic Variables on the Amyloid-β Deposit Detection/Classification Model

The Amyloid-β deposit detection with subsequent plaque classification outputs counts for cored and diffuse Amyloid-β plaques. .The module acquires these counts by detecting and then classifying all deposits located in the WSI. A comparison of these predictions for a single case can be seen in Figure 3. We observe a degree of disagreement in prediction between the different datasets for both diffuse as well as cored plaques. Figure 3 shows a case example with heap maps and accompanying quantitative results for plaque counts in background, GM, and WM, and Figure 4 is a graphical representation of the quantitative results for cored and diffuse plaque counts and GM/WM ratios across all cases based on the pre-analytic variable.

By acquiring the quantitative results for cored /diffuse plaque counts and applying the ANOVA test, we can test whether the pre-analytical variables affect the target outcome (deposit counts). Table 1 shows magnification and scanner type are two pre-analytical variables with the most effect on our DL predictions. Results from Table 1 show similar effect observed in the case of the segmentation model, where magnification and scanner (GT450) are the pre-analytical variables with the most effect on DL predictions.

II. Effects of Pre-Analytic Variables on the WM/GM Segmentation Model

The WM/GM segmentation module yields WM/GM ratio as a quantitative measure that can be used in statistical analysis. We plotted the WM/GM ratios for the different datasets evaluated in Figure 4. When applying ANOVA test to those values, we can check whether the pre-analytical variables affect the target (WM/GM ratio). Table 1 summarizes our results; both magnification and scanner type (GT450) have significant effects on DL prediction outcomes.

The WM/GM segmentation map also outputs a heatmap of the segmented WSI, in this heatmap we have GM predictions denoted as cyan, WM as yellow and background as black (Figure 3, supplemental figures 2, 4). This method of visualizing the results is a better indicator of stable performance, as that is the final product to be analyzed by the expert, as well as the map to be used for calculation of densities of deposits and structures seen in the WM/GM. As seen in Figure 4, changing scanners and magnification has an effect on our model's predictions. For case 7, when comparing the results from AT2-20X, Axio-Z1-40X-75, and Axio-Z1-40X-45, there are prediction disagreements between GM and WM close to the boundaries of GM and background (supplemental figure 2).

III. Saliency Maps

When analyzing heatmaps and quantitative scores acquired from the two DL frameworks, we can assess how the pre-analytical variables affect the outputs. However, this information only tells us how the final output was affected, but the effect on the prediction process of the DL frameworks is still unknown. Saliency maps allow us to tap into the blackbox nature of DL models and learn a bit about their prediction process, more specifically, how much the different locations in each image contributed to the final output. We employed Class Activation Mapping (CAM) (25), Grad-CAM (26) and Grad-CAM++ (27) as methods to acquire the saliency maps. By analyzing the saliency maps generated from each 256x256 patch, we can observe which areas of each patch contributed most to the final model output and how these areas may differ according to the pre-analytical variables. That is especially relevant for the WM/GM segmentation DL model, as there is no obvious single structure linked to the predictions such as an Amyloid-β plaque classification, as it relies on features such as texture of the tissue, as shown previously (23).

The Grad-CAM presented in Figure 5 shows that despite the prediction outcome remaining constant as GM in all the cases shown, the areas that led the WM/GM frameworks to reach that conclusion were different. The Grad-CAM++ displays less differences, pointing towards a higher level of agreement that occurs when taking in consideration a more complex interpretability framework. The same effect is observed when the agreed predictions are WM, as seen in Figure 5.

We are also able to see in supplemental figure 3 the difference remains when the final output disagrees. This patch is taken from the patch with high WM-GM prediction disagreement observed between AT2-20X and Axio-Z1 datasets in case 7 (seen in Supplemental Figure 2). Despite a level of overlap in the saliency map, the AT2-20X CAM covers a wider area than its Axio-Z1 counterparts.

Discussion

Recent studies utilizing ML/DL in pathology have been successful in displaying high prediction performance (7, 8, 9, 23). However, due to the blackbox nature of trained DL models, rigorous testing is needed since there is no guarantee a model trained with a set

of data may have the same performance when applied to data with different pre-analytic variables. Furthermore, as other studies have shown feeding WSIs with different pre-analytical variables may degrade performance (17), we must extend the rigorous testing to account for such differences. However, studies that include this generalizability test are limited (17-20).

Our study's AT2-20X dataset constitutes a fair baseline: it shares the same scanner, magnification, brain region, storage format, stain, and compression standard as the training data. Therefore, when performing tests on this subset of data, we expected the model to achieve the most accurate performance in both DL modules as there is no variance in pre-analytical variables. In the current publication, we have utilized the AT2-20X in the training set, so we set it as the gold standard of performance for all other datasets.. By evaluating the DL frameworks on AT2-20X and then evaluating on other datasets with different pre-analytical variables, we investigated the effect pre-analytical variables on DL generalizability in WSIs and observed some level disagreement in both DL frameworks.

In this pilot study, we demonstrate the detection of Amyloid-β plaques in brain WSI trained on 20X AT2 slides is affected by WSI magnification (40X) and GT450 scanner. We see a similar effect for both diffuse, and cored Amyloid-β plaques. In addition to the statistical analysis result, we can reliably observe an overall effect of pre-analytical variables on Amyloid-β plaque counts when observing heatmaps and counts per area.

We also observed performance differences when our generalizability test was applied to WM/GM segmentation task. Our results revealed WSI magnification (40X) and scanner type (GT450) have an impact on predicted WM/GM ratio. Our observation of WM/GM segmentation heatmaps and saliency maps also displays unstable WM/GM predictions when applied to WSIs from GT450 scanner (see figure 5 for example). Some outlier WM/GM predicted heatmaps can be observed in Zeiss scanners having stark differences when compared to the AT2-20X heatmap (see supplemental figure 4).

The outlier performance from GT450 expanded to all cases employed in this study. When analyzing the overall scan from the GT450 in comparison to other scanners, subjective observations denoted an increase in brightness and white tones. We hypothesized the different standard color profile applied to the scan (i.e. ICC profile) is responsible for the difference observed. Since scanning was done with default parameters, the difference in color profile may extend to the software version employed at the time of the scan. Studies argue a normalization step is required to match performance between scanners or different scanning protocols (31). In future works, we aim to examine how color normalization may alter results within the pre-analytic variable realm (32,33,34). Preliminary experiments show Reinhard normalization (32) as a suitable intervention to address GT450 performance differences (see supplemental figure 5).

Although this study contributes to the field, there are some limitations to consider. First there is the misalignment of digitalized tissue caused by different scanners' field of view

Due to such misalignment, we could not perfectly overlap the heatmap predictions. Such limitation prevented us from using our WM/GM annotated ground truth to reliably calculate Intersection over Union (IoU) or DICE coefficient between our WSIs digitized from the same slides. Both DICE coefficient and IoU have been used to compare ground truth and predictions on a pixel-by-pixel basis (35-36). This misaligning prevented automated tile comparison, as a human observer was required to fine-tune registration for each individual area compared. There is also the use of only 20x and 40x magnification; additional works with other magnifications such as 10x and 5x may be advantageous as file sizes may be smaller and easier to process. Third, our study examined only a limited number of cases from a single brain bank. Due to the large file size of WSIs, especially at 40X magnitude, it becomes a time-consuming task to generate predictions for both WM/GM segmentation and Amyloid-β deposit assessment, approximately 6 hours per 20X slide when employing a NVIDIA Tesla T4 GPU. In our study, we processed a total of 65 slides, which account for almost 400 hours of GPU use. Lastly, we utilized the AT2-20X as the gold standard to conduct comparisons in the current study. To our knowledge, although checklists for machine learning algorithms in medical imaging have been proposed (34) there are no gold standards for pre-analytic variables for digital pathology when conducting machine learning algorithms. The choice of using AT2-20X as gold standard is due to data of same pre-analytical variables being employed in training. This choice best matches the recommendations of item 7, regarding data sources, in previous works (37) as test data from AT2-20X matches the trained model best. Unlike other medical imaging fields such as Radiology that have standard file formats, there are many options given the vast array of available slide scanners and associated settings in the WSI realm. This study highlights the importance of denoting scanner types, magnifications, as well as compression rates when conducting such workflows.

Generalizability is a crucial challenge for deploying DL in real life pathology problems. Currently in the field there are studies seeking to perform Domain Adaptation (DA) techniques to address generalizability from diverse pre-analytical variables in ML frameworks (28-30). These efforts are important to advance the generalizability of frameworks in the field and address the unwanted effects we observe when varying pre-analytical variables. Unlike normalization, DA does not need any additional preprocessing steps for generalization to many different scanners. The application of these DA techniques has also been shown in the WSI domain (30) and would be a great candidate to address the performance difference observed in this study.

Acknowledgments

The authors thank the families and participants of the University of California Davis Alzheimer's Disease Research Centers (ADRC) for their generous donations as well as ADRC staff and faculty for their contributions. Resources for this study were funded in part by grants from the National Institute on Aging of the National Institutes of Health under Award Numbers R01AG062517, P30AG072972, and R01AG056519, and a research grant from the California Department Of Public Health (19-10611) with partial funding from the 2019 California Budget Act. The views and opinions expressed in this manuscript are those of the author and do not necessarily reflect the official policy or position of any public health agency of California or of the United States government. We

also thank the UC Davis Health Department of Pathology and Laboratory medicine as well as the laboratory of Dr. Alexander “Sandy” Borowsky for the use of digital slide scanners.

References

1. Al-Janabi S, Huisman A, Van Diest PJ. Digital pathology: current status and future perspectives. *Histopathology* 2012;61:1-9

2. Bankhead P, Loughrey MB, Fernandez JA, et al. QuPath: Open source software for digital pathology image analysis. *Sci Rep* 2017;7:16878

3. Dugger BN, Dickson DW. Pathology of Neurodegenerative Diseases. *Cold Spring Harb Perspect Biol* 2017;9

4. Shakir MN, Dugger BN. Advances in Deep Neuropathological Phenotyping of Alzheimer Disease: Past, Present, and Future. *J Neuropathol Exp Neurol* 2022;81:2-15

5. McKenzie AT, Marx GA, Koenigsberg D, et al. Interpretable deep learning of myelin histopathology in age-related cognitive impairment. *Acta Neuropathologica Communications* 10.1 2022;1-17.

6. Vizcarra JC, Gearing M, Keiser MJ, et al. Validation of machine learning models to detect amyloid pathologies across institutions. *Acta neuropathologica communications* 8.1 2020;1-13.

7. Tang Z, Chuang KV, DeCarli C, et al. Interpretable classification of Alzheimer's disease pathologies with a convolutional neural network pipeline. *Nat Commun* 2019;10:2173

8. Lai Z, Guo R, Xu W, et al. Automated grey and white matter segmentation in digitized aβ human brain tissue slide images. *2020 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*. IEEE, 2020.

9. Litjens G, Kooi T, Bejnordi B, et al. A survey on deep learning in medical image analysis. *Medical image analysis* 42 2017;60-88.

10. Hekler A, Utikal JS, Enk AH, et al. Deep learning outperformed 11 pathologists in the classification of histopathological melanoma images. *European Journal of Cancer* 118 2019;91-96.

11. Aresta G, Araujo T, Kwok S, et al. Bach: Grand challenge on breast cancer histology images. *Medical image analysis* 56 2019;122-139.

12. Hsu WW, Guo JM, Pei L, et al. A weakly supervised deep learning-based method for glioma subtype classification using WSI and mpMRIs. *Scientific Reports* 12.1 2022;1-12.

13. Kiani A, Uyumazturk B, Rajpurkar P, et al. Impact of a deep learning assistant on the histopathologic classification of liver cancer. *NPJ digital medicine* 3.1 2020;1-8.

14. Balkenhol MC, Tellez D, Vreuls W, et al. Deep learning assisted mitotic counting for breast cancer. *Laboratory investigation* 99.11 2019;1596-1606.

15. LeCun Y, Boser B, Denker J, et al. Handwritten digit recognition with a back-propagation network. *Advances in neural information processing systems* 1989;2

16. Vali-Betts E, Krause KJ, Dubrovsky A, et al. Effects of image quantity and image source variation on machine learning histology differential diagnosis models. *Journal of Pathology Informatics* 12.1 2021;5.

17. Jang HJ, Song IH, Lee SH. Generalizability of deep learning system for the pathologic diagnosis of various cancers. *Applied Sciences* 11.2 2021;808.

18. Yan W, Huang L, Xia L, et al. MRI manufacturer shift and adaptation: increasing the generalizability of deep learning segmentation for MR images acquired with different scanners. *Radiology: Artificial Intelligence* 2.4 2020.

19. Sathitratanacheewin S, Sunanta P, Pongpirul K. Deep learning for automated classification of tuberculosis-related chest X-Ray: dataset distribution shift limits diagnostic performance generalizability. *Heliyon* 6.8 2020; e04614.

20. Jones AD, Graff JP, Darrow M, et al. Impact of pre-analytical variables on deep learning accuracy in histopathology. *Histopathology* 75.1 2019;39-53.

21. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* 2014.

22. He K, Zhang X, Ren S, et al. Deep residual learning for image recognition. *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016.

23. Lai Z, Oliveira LC, Guo R, et al. BrainSec: Automated Brain Tissue Segmentation Pipeline for Scalable Neuropathological Analysis. *IEEE Access* 10 2022;49064-49079.

24. Jiang J, Larson NB, Prodduturi N, et al. Robust hierarchical density estimation and regression for re-stained histological whole slide image co-registration. *Plos one* 14.7 2019; e0220074.

25. Zhou B, Khosla A, Lapedriza A, et al. Learning deep features for discriminative localization *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016.

26. Selvaraju RR, Cogswell M, Das A, et al. Grad-cam: Visual explanations from deep networks via gradient-based localization. *Proceedings of the IEEE international conference on computer vision*. 2017.

27. Chattopadhay A, Sarkar A, Howlader P, et al. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. *2018 IEEE winter conference on applications of computer vision (WACV)*. IEEE, 2018.

28. Breen J, Zucker K, Orsi NM, et al. Assessing domain adaptation techniques for mitosis detection in multi-scanner breast cancer histopathology images. *International*

Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, Cham, 2021.

29. Aviles J, Talou GD, Camara O, et al. Domain Adaptation for Automatic Aorta Segmentation of 4D Flow Magnetic Resonance Imaging Data from Multiple Vendor Scanners. *International Conference on Functional Imaging and Modeling of the Heart*. Springer, Cham, 2021.

30. Panfilov E, Tiulpin A, Klein S, et al. Improving robustness of deep learning based knee mri segmentation: Mixup and adversarial domain adaptation. *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*. 2019.

31. Khan AM, Rajpoot N, Treanor D, et al. A nonlinear mapping approach to stain normalization in digital histopathology images using image-specific color deconvolution. *IEEE transactions on Biomedical Engineering* 61.6 2014;1729-1738.

32. Reinhard E, Adhikhmin M, Gooch B, et al. Color transfer between images. *IEEE Computer graphics and applications* 21.5 2001;34-41.

33. Roy S, Panda S, Jangid M. Modified Reinhard Algorithm for Color Normalization of Colorectal Cancer Histopathology Images. *2021 29th European Signal Processing Conference (EUSIPCO)*. IEEE, 2021.

34. Vahadane A, Peng T, Sethi A, et al. Structure-preserving color normalization and sparse stain separation for histological images *IEEE transactions on medical imaging* 35.8 2016; 1962-1971.

35. Rahman MA, Wang Y. Optimizing intersection-over-union in deep neural networks for image segmentation. InInternational symposium on visual computing *International symposium on visual computing*. Springer, Cham, 2016.

36. Zou KH, Warfield SK, Bharatha A, et al. Statistical validation of image segmentation quality based on a spatial overlap index1: scientific reports *Academic radiology* 11.2 2004;178-189.

37. Mongan J, Moy L, Kahn Jr CE. Checklist for artificial intelligence in medical imaging (CLAIM): a guide for authors and reviewers. *Radiology. Artificial Intelligence* 2.2 2020.

Figure legends

Figure 1: Schematic representation of pre-analytical variables evaluated. Variables linked by arrows are nested, for example, all Zeiss Axioscan Z1 data employed will be of CZI storage format.

Figure 2: CNN-based DL pipeline employed in this study. The approach for prediction is patch-based, therefore WSI must be patched prior to analysis. The different blue and green segmented circled on the grided images in the top figure (light orange box) panel refer to the different resolutions (1536 x 1536, and 256 x 256, respectively) patched by the framework. There are two DL modules responsible for predictions (lower figure panel-light blue box), a WM/GM segmentation and an Amyloid-β deposit detection with subsequent classification module which operate on the 256x256 pixel resolution. For heatmaps in the bottom left corner of the figure, white matter (WM) is represented in yellow, grey matter (GM) as cyan, background as plaque and plaques at orange. (figure adapted from (23)).

Figure 3: Schematic of heatmaps of GM/WM segmentation and plaques counts for case 4. Top panel- GM/WM cored plaque heatmap (left) and counts by area based on select pre-analytic variables (right). Bottom panel, GM/WM diffuse plaque heatmap (left) and counts by area based on select pre-analytic variables (right). A zoomed-in area (not the WSI) of case 4 was chosen to aid in visualization. For heatmaps, plaques are depicted in orange, background as black, White matter (WM) as yellow, and Grey Matter (GM) as cyan.

Figure 4: WM/GM ratio, cored, and diffuse plaque (A), Cored plaque (B) counts and WM/GM ratio (C) for each case by pre-analytic variable. Cases with none/low likelihood AD (5, 6, 8, 10, 11, 14) typically had low numbers of core plaques, while cases with high likelihood AD (1, 3, 4, 12, 13) had higher counts. More information on the demographics of cases located in Supplemental Table 2. Further details on case 6 for GM/WM ratio is located within supplemental figure 4.

Figure 5: Grad-CAM and Grad-CAM++ of agreed predictions of Grey matter (GM) and White matter (WM). All datasets agreed on the tile's prediction and got the correct prediction.

Supplemental Figure 1: Example of differential alignments, sizes, and aspect ratios using the WSI from Case 1. Comparing to the AT2-20x, the GT450-40X has changes in pixel intensities, theorized to be caused by different color profile (i.e. ICC profile); the Axio-Z1-40X-75 had cropping of the tissue section with larger aspect ratios theorized to be caused by automatic tissue detection.

Supplemental Figure 2: Heatmap of WM/GM predictions for the different datasets equivalent of case 7. Original AT2-20X WSI included for reference.

Supplemental Figure 3: CAM of disagreed predictions of GM. The tiles were taken from case 7, which has its WM/GM predictions present in Supplemental Figure 2. Tiles and saliency maps from the original AT2-20X dataset, which correctly predicted GM as GM are included. Tiles and saliency maps from Axio-Z1-40X-75 and Axio-Z1-40X-45 are included, the output for these two was WM, an incorrect prediction.

Supplemental Figure 4: Case 6 WSI from AT2-20X and heatmap predictions from Axio-Z1-40X-75, Axio-Z1-40X-45, and AT2-20X.

Supplemental Figure 5: Plots of cored and diffuse plaque counts on GT450 before and after Reinhard (32) normalization was applied (using a AT2 WSI as reference slide). We compare results withAT2-20X for reference.

Table 1

Statistical Analysis Values

	Axio-Z1-40X-45 Vs. Axio-Z1-40X-75	**AT2-20X Vs. Axio-Z1-40X-75**	**AT2-20X Vs. GT450-40X**	**AT2-20X Vs. AT2-40X**
Cored Count	0.4024	0.0738	0.0078	0.0160
Diffuse Count	0.2272	0.4290	0.0073	0.0705
WM/GM Ratio	0.0853	0.2475	0.0005	0.0013

Table 1: P-values for ANOVA tests

Figure 1

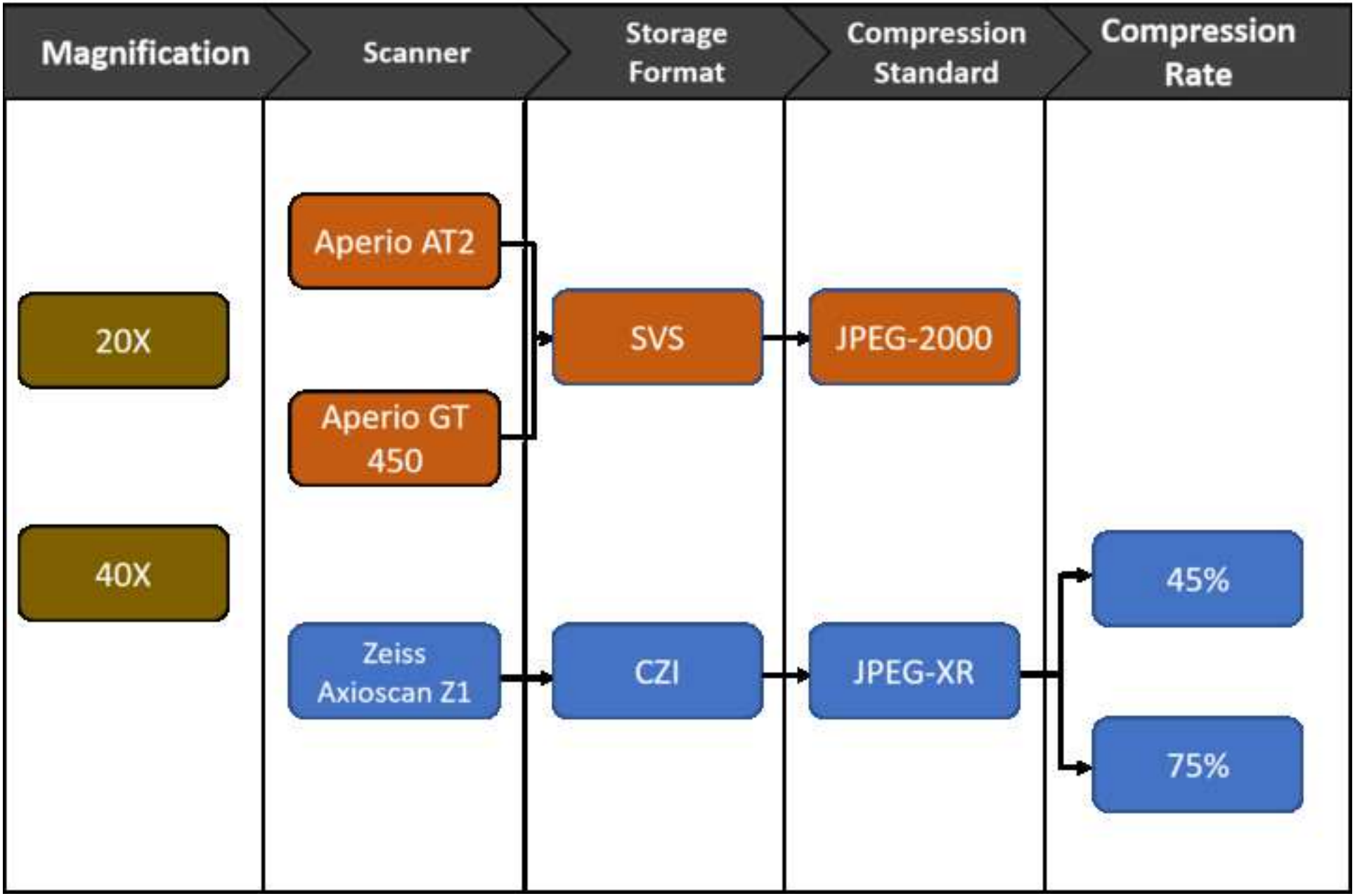


Fig 2

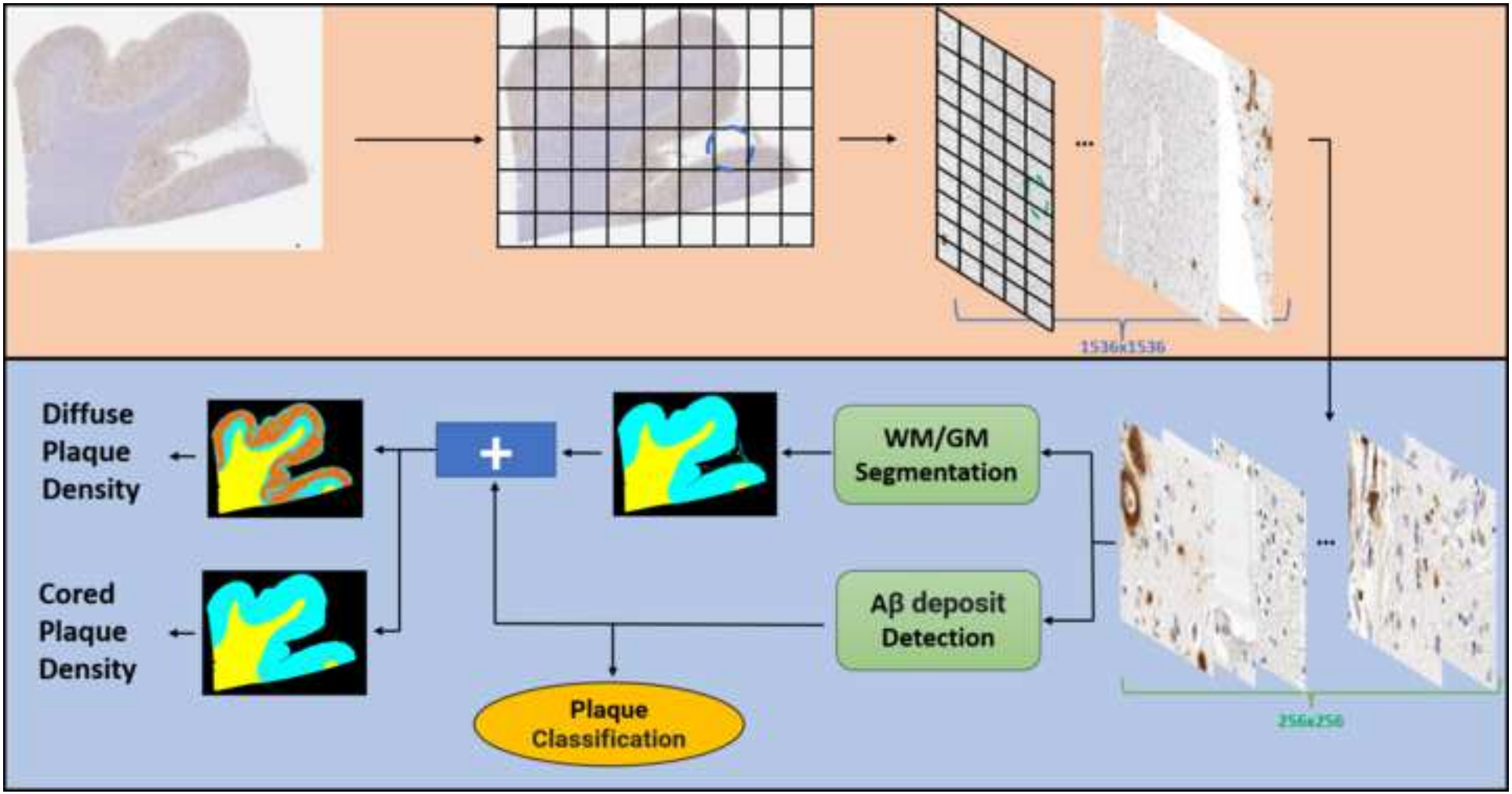


Fig 3

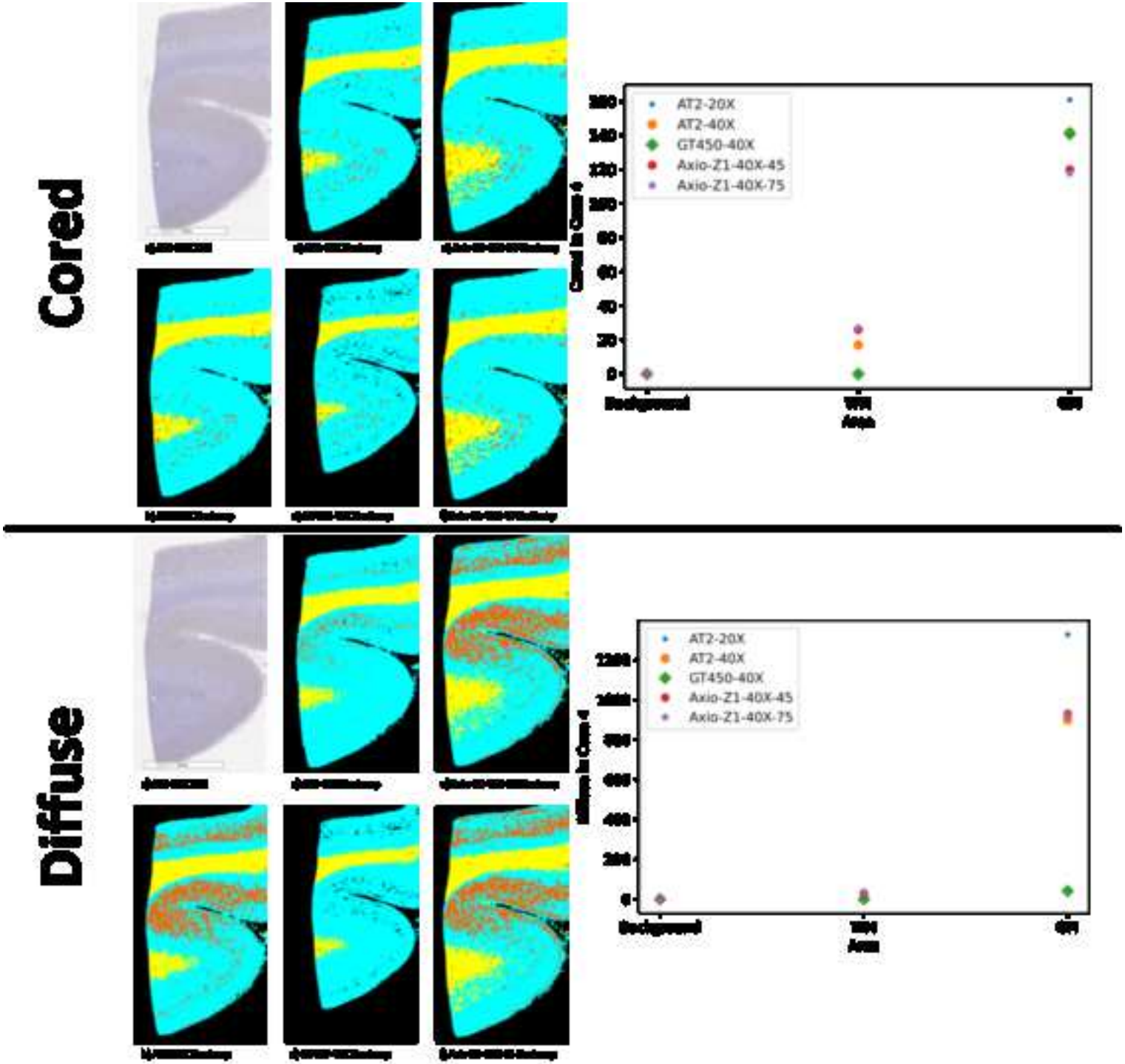
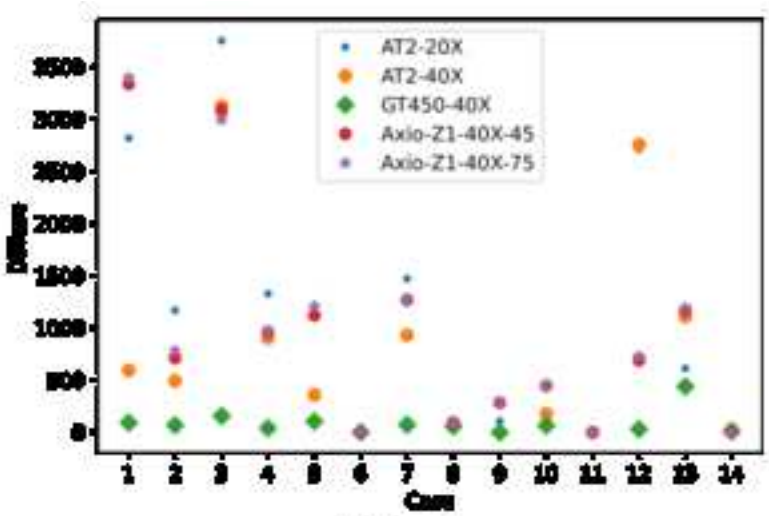
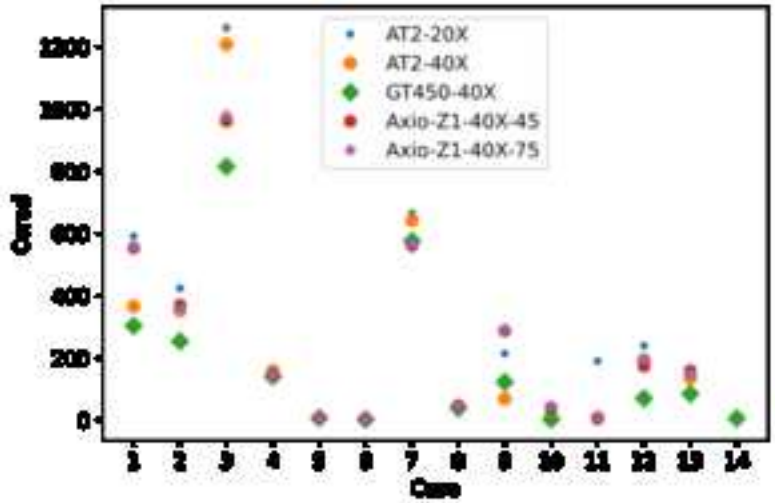


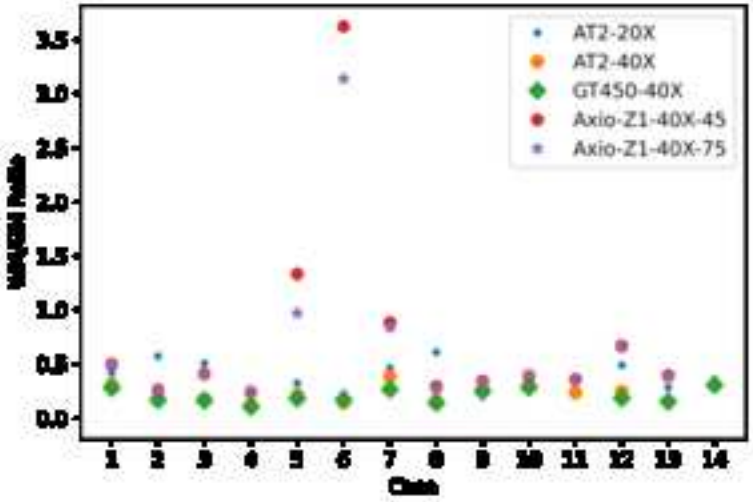
Fig 4



a) Diffuse Count



b) Cored Count



c) WM/GM Ratio Plot

Fig 5