AMYLOID-BETA PLAQUE QUANTIFICATION AND ANALYSIS

Thesis report

to obtain the degree of Master of Science
at the Delft University of Technology,
to be defended publicly on Monday September 4, 2023 at 09:00 AM.

by

Chiel DE VRIES

Student number: 4571983
Project duration: November 15, 2022 – September 4, 2023
Thesis committee: Prof. dr. ir. M. J. T. Reinders,   TU Delft, supervisor
                    Dr. ir. X. Zhang,   TU Delft
                    Dr. ir. T. Höllt,   TU Delft
                    Dr. J. J. M. Hoozemans,   VU Amsterdam
                    S. Rhode,   VU Amsterdam

TUDelft
Delft
University of
Technology

ABSTRACT

Alzheimer's disease (AD) is becoming more prevalent as the world population gets older. The formation of Amyloid-beta ($A\beta$) plaques is one of the pathologies related to AD. Recent work has shown that the $A\beta$ load in brain tissue has a negative correlation with cognitive performance in cognitively healthy centenarians. This work aims to expand this research by investigating whether the types of $A\beta$ plaque present are linked to cognition and by comparing the types of plaques in the centenarian cohort with an AD cohort. For this task, a system is developed that can identify $A\beta$ plaques in images of brain tissue. It first automatically segments the grey matter using a fine-tuned U-net. Then the plaques are located using traditional image processing techniques. Lastly, shape and size features are extracted from the plaque in addition to a feature vector made by a pre-trained AlexNet. K-means clustering is used on AlexNet features to find categories for the plaques. The clustering approach failed to yield good results. However, the area and roundness are differently distributed between the AD and centenarian cohorts, but the differences are small. Correlations have been found between the area and roundness of plaques in the occipital pole and cognitive performance in centenarians. They indicate that cognitively stronger individuals have smaller and less round $A\beta$ plaques in their brains. More research is necessary to reveal the true extent of the impact of plaque types on cognition.

CONTENTS

1

INTRODUCTION

As the global population ages, the prevalence of age-dependent neurodegenerative diseases is on the rise. However, not all elderly people develop such diseases. The 100+ study [1], based at the Amsterdam University Medical Centers (AUMC), investigates this group of people. They visit self-reported cognitively healthy centenarians, people of more than 100 years of age, for a neuropsychological examination. In addition, they collect genetic data and participants have the option to donate their brains after they pass away. This has proven to be an invaluable source of data on its own and as a means of comparison with cases affected by neurodegenerative diseases.

The 100+ study focuses mainly on Alzheimer's disease (AD). It is one of the most common neurodegenerative diseases, accounting for 60% to 80% of cases of dementia in the US [2]. AD is characterized by several pathologies [9], of which the formation of amyloid-beta (Aβ) plaques is the focus of this work. Aβ plaques are protein deposits that form in the extracellular space in brain tissue, mainly in the grey matter.

Aβ can be deposited in various forms. The primary differentiation is made between dense and diffuse depositions, but they can be separated into more precise categories. Currently, there is no standardised list of types available [33]. However, the following are mentioned in the literature.

A *Diffuse plaque* (also known as a primitive or immature plaque) is a gathering of many small deposits of Aβ scattered around the tissue (Figure 1.1a). A *Cored plaque* (also known as a 'mature' or 'classical' plaque) is a plaque with a dense core surrounded by an empty space which is followed by a corona of diffuse Aβ (Figure 1.1b). A *Compact plaque* (also known as a burned-out plaque) is similar to a cored plaque, but without the corona (Figure 1.1c). As their alternative names suggest, plaques are thought to start out as diffuse, then progress to cored plaques and end as compact plaques [36]. However, this progression remains speculative [13]. Furthermore, a *Coarse-grained plaque* is a plaque with multiple small cores and with pores that do not contain Aβ (Figure 1.1d). *Cerebral amyloid angiopathy (CAA)* is Aβ that accumulates in the vascular wall forming a ring-like structure. Depending on the angle at which the blood vessel is dissected, it can appear elongated. (Figure 1.1e) Lastly, *Subpial aggregations* are diffuse depositions
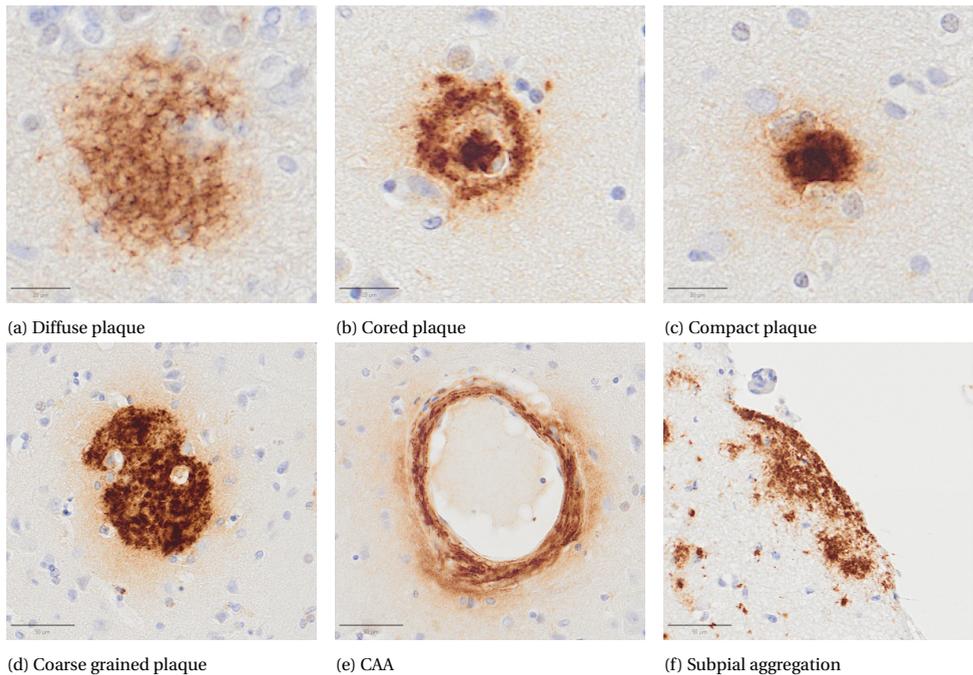
1

(a) Diffuse plaque

(b) Cored plaque

(c) Compact plaque

(d) Coarse grained plaque

(e) CAA

(f) Subpial aggregation

Figure 1.1: Examples of various different plaque types. Scale bars are included in the bottom left of each image. (a)-(c) Bar = 20 µm. (d)-(f) Bar = 50 µm.

of Aβ that have a band-like structure and form underneath the pia mater. (Figure 1.1f)

This is not a comprehensive list of all plaque morphologies, but it illustrates several ways the Aβ may be deposited. Furthermore, not all plaque types occur at the same rate and they are not always easily identifiable. Trained experts may have different opinions on how a plaque should be categorised. This makes it harder to quantify and analyse the causes and effects of the plaques.

However, advancements in imaging techniques and tools allow Aβ pathology to be analysed more and more digitally [24]. A high-resolution image of a slide can be made, called a whole slide image (WSI). See figure 2.1 for an example of a WSI. Initially, they were utilised to archive slides efficiently, but the technique is increasingly used to help with the analysis of slides as well [12]. This opens the door for the application of computer vision techniques to step in. Large amounts of data can be quantitatively analysed much more quickly than by hand.

Stephen et al. [31] used a random forest classifier to detect extracellular dense and diffuse Aβ plaques as well as intracellular Aβ precursor protein deposits. They showed that the more advanced AD cases in their cohort have a higher proportion of diffuse plaques when compared to less advanced cases.

Other works use a convolutional neural network (CNN) to build a classification model that predicts whether the input image contains any of three types of Aβ plaques: cored, diffuse and/or CAA [14, 32]. Using a sliding window approach or by splitting the WSI into

tiles, they analyse a full WSI to predict the location and type of plaque.

Lai et al. [21] added a segmentation model to this pipeline that can predict whether an image is part of the grey matter, white matter or background. With this system, they were able to find how plaques and plaque types are distributed between the two types of brain tissue. Their results were consistent with previous findings [18], showing that most plaques appear in the grey matter and, to a much lesser extent, they also appear in the white matter.

An unpublished study from the AUMC has designed a method to calculate the Aβ load in a WSI [15]. The Aβ load is the percentage of the area of the grey matter that is covered in Aβ deposits. The implementation was made using the pixel classifier available in Qupath [5], an open-source bioimage analysis tool. Using the data collected by the 100+ study they were able to find significant correlations between cognitive performance and Aβ load in centenarians. In addition, a comparison was made between the 100+ cohort and a set of slides from AD cases collected by the AUMC. They found that the load in AD cases was generally much higher than in 100+ cases. However, there were slides in the 100+ cohort that had similar or higher Aβ loads than AD cases. How could it be that these cases were still considered cognitively healthy?

This work is a continuation of this unpublished study by Hernández [15]. After finding correlations between the Aβ load and cognitive performance, the question arose whether the occurrence of specific plaque types might be related to cognitive performance as well. Furthermore, are the Aβ plaques types in AD cases distributed differently than the cognitively healthy centenarians in the 100+ cohort? To answer these questions, a method must be developed to quantify Aβ plaque types in WSIs of brain tissue.

This work aims to design and build a system that performs this task. It should be able to automatically locate Aβ plaques in the grey matter of the brain tissue. Then it should be able to extract information about what kind of plaque it is. To attain this goal, the system is built in three steps.

- The first step is *segmentation* (section 2.2). In this step, the grey matter is segmented from the WSI. The Aβ plaques are mostly located in the grey matter. Therefore checking the white matter and background of the WSI for Aβ is not necessary. Since WSIs are very large, this can significantly reduce the amount of data that needs to be analysed, improving the overall computation time of the system. Additionally, there are often areas in the white matter that did pick up some of the staining dye, even though no Aβ is present. Thus the segmentation can reduce the number of false positive plaque detections. While part of the dataset has the grey matter annotated (section 2.1), not all of it does. Therefore, an automated grey matter segmentation algorithm is required. This algorithm would not only be useful for this work, but also for further research using a different, unannotated dataset.

- The second step is *localisation* (section 2.3). The goal of this step is to analyse the grey matter to find individual plaques. It should locate them, so they can be cropped from the WSI and analysed further to collect their characteristics. If this is done automatically it means that large amounts of WSIs can be scanned quickly, making the quantitative analysis on plaques feasible.

1

- The third step is *categorisation* (section 2.4). Here the individual plaques are analysed to extract information from them so they might be categorised and described accordingly. Since there is no exhaustive set of plaque types available, the goal is to let a clustering algorithm find appropriate categories. A clustering model learns unsupervised, therefore the plaques do not need to be annotated, saving valuable time for the trained experts. The categories found with this method may shine a new light on the types of Aβ plaques that occur in brain tissue.

2

METHODS

2.1. DATASETS

This work makes use of two datasets provided by the Amsterdam University Medical Centers. They are not publicly available because they contain sensitive personal information. A small third set was created specifically for this work. An overview of the two large datasets can be found in table 2.1.

2.1.1. THE 100+ DATASET

The first dataset consists of 462 slides collected from 93 different individuals. The individuals in this dataset donated their brains as part of the 100+ study [1]. Samples of brain tissue were taken from six cortical brain regions. Each sample is 8 μm thick and immunohistochemically stained for the Aβ protein. The samples were scanned using the Olympus VS2000 slide scanner to create a WSI for further analysis. Each WSI has 3 colour channels (red-green-blue) and they have an extremely large resolution to ensure all cellular details are retained. Each pixel spans 0.274 μm, which means the full image can have 50.000 to 100.000 pixels per axis depending on the size of the tissue sample. After the slides were digitised, the grey matter was annotated by an expert for each WSI.

In addition to the WSIs, the dataset contains the scores from a set of neuropsychological tests that were performed annually before brain donation. The tests that are included can be found in table 2.2. The dataset also contains which allele each individual has for the APOE gene. This gene is associated with an increased or decreased risk of developing AD [3]. Lastly, the Aβ load values from the unpublished study by Hernández [15] are available for each WSI.

2.1.2. THE AD DATASET

The second dataset contains 116 slides collected from 29 different individuals. The individuals in this dataset were diagnosed with AD before they donated their brains. The slides were prepared and digitised in the same manner as the 100+ dataset. The tissue samples in this group were only taken from 4 cortical brain regions. The dataset does not
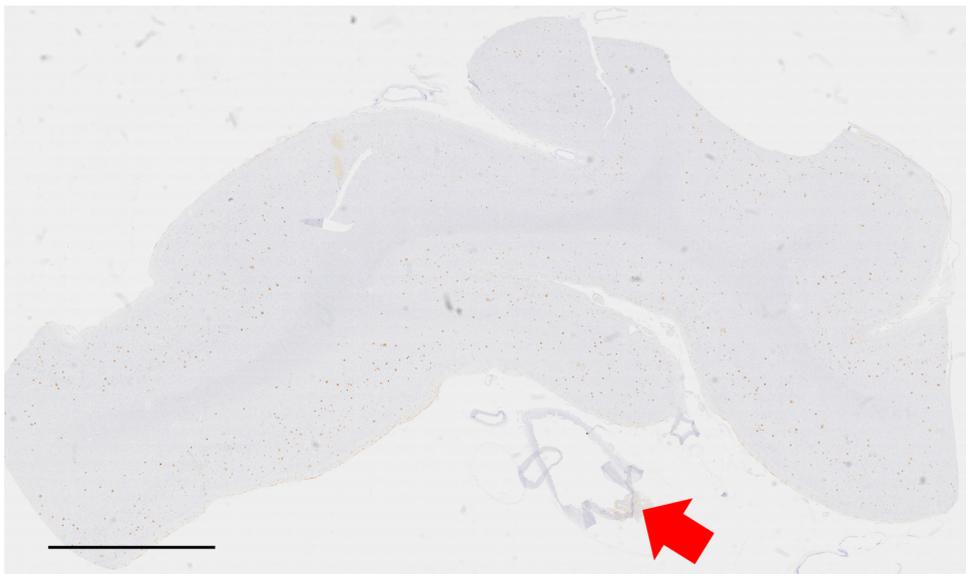
5

2

Figure 2.1: An example of a WSI of a sample of brain tissue. The sample is immunohistochemically stained such that the cell nuclei are blue and the Aβ plaques are brown. The background is white and the brain tissue is blue-grey. The brain tissue can be subdivided into the grey matter and the white matter. Counterintuitively, the white matter is slightly darker with this staining, due to the increased amount of nuclei that are present there. The image covers an area of approximately 28 by 17 mm, which corresponds to approximately 102,000 by 62,000 pixels.

Table 2.1: Statistics on the 100+ and AD datasets. It shows the number of slides and individuals in each set, as well as the number of slides from each brain region in the set. For the image height and width, the average and standard deviation over the set are shown.

	100+	**AD**
Number of slides	462	116
Number of individuals	93	29
Middle frontal gyrus	93	29
Middle temporal gyrus	52	0
Temporal pole	90	29
Inferior parietal lobe	93	29
Parietal lobe	40	0
Occipital pole	94	29
Image height in pixels	73116 ± 8502	72956 ± 8990
Image width in pixels	96012 ± 10410	100863 ± 15190

Table 2.2: The set of neuropsychological tests performed on participants of the 100+ study.

Test	Description
Mini Mental State Examination (MMSE)	A 30-point questionnaire that is filled in by the participant. It measures cognitive impairment.
Digit Span-backward	The participant listens to a series of digits and has to repeat them backwards.
Key Search	The participant needs to draw a path through a field they would walk to search for an imagined set of lost keys.
Clock Drawing Test	The participant is asked to draw the numbers on the face of a clock. Then they are asked to draw a specific time.

have the neuropsychological test scores and grey matter annotation, but it does have the APOE allele for each individual and the Aβ load for each WSI as calculated by Hernández [15].

2.1.3. THE PLAQUE DETECTION DATASET
The third dataset is much smaller than the previous two. The set was created to evaluate the localisation step. It contains twelve images of 4096 by 4096 pixels, roughly 1.1 by 1.1 mm. These images are hand-picked from the 100+ dataset and chosen to be as varied as possible, because it is important that the system can recognise plaques in different conditions. The first variable is the intensity of the staining, the second is the size of the plaques, and the third is the number of plaques in the patch. An expert annotated 1686 plaques in the twelve images.

2.2. SEGMENTATION
The first step of this work is to segment the grey matter from the WSIs. This is necessary because the Aβ plaques accumulate in that part of the brain. Segmenting the grey matter, therefore, helps to lower the amount of data that needs to be analysed downstream, thus lowering the overall computation time. Furthermore, a segmentation algorithm could be used by itself as well, as part of a pathologist's toolbox. The segmentation algorithm consists of a deep neural network (DNN) trained on patches from the 100+ dataset. That dataset has human-annotated ground truth labels for the grey matter, allowing it to be used for training.

2.2.1. MODEL
The model used by this work is a pre-trained U-net [29] architecture implemented by Buda et al. [6]. U-net is a fully convolutional network (FCN) that features an encoder-decoder structure, as is visible in figure 2.3. The contracting path consists of four blocks of two 3x3 convolutions with rectified linear unit (ReLU) activation functions and a max pooling layer to downsample the input for the next block. In addition, the number of channels is doubled with each block. At the bottom, there are two more 3x3 convolu-

2

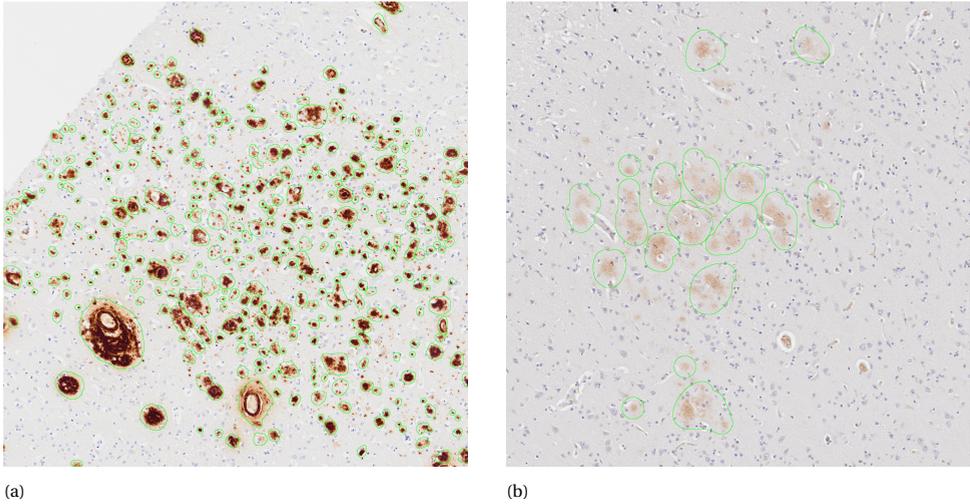(a)   (b)

Figure 2.2: Two of the twelve images from the plaque detection set. Their width and height are 4096 pixels, which corresponds to approximately 1.1mm. The annotations are shown in green. (a) has a dark staining with small plaques. (b) has a lighter staining and contains fewer but larger plaques.
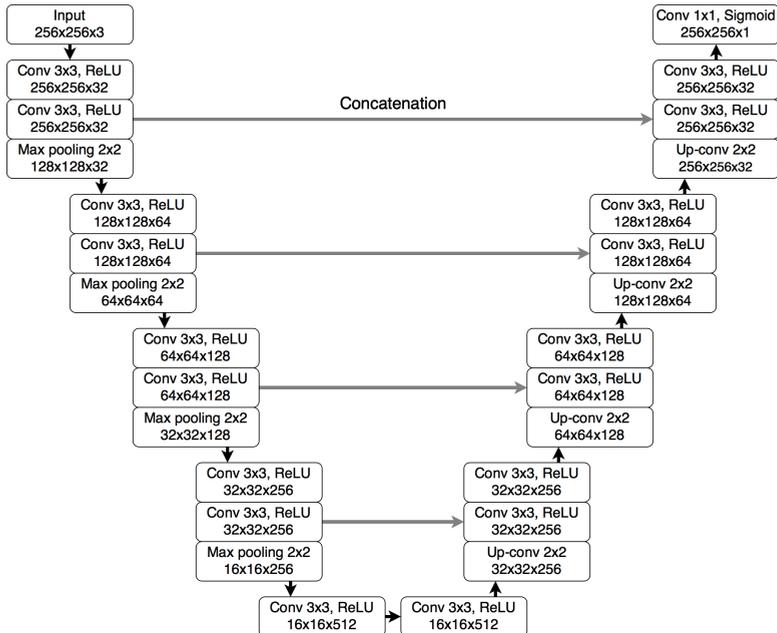

Figure 2.3: The U-net architecture as implemented by Buda et al. [6]. This visual shows an example input image of 256 by 256 pixels and 3 colour channels.

tions with ReLU activations. This is followed by the expanding path consisting of four blocks of an up convolution and two 3x3 convolutions with ReLU. Each block on the same level is connected by a skip connection that concatenates the output of each contracting block to the input of the expanding block on the channel axis. This is done to retain information on each detail level in the image. The final layer reduces the number of channels to one by using a 1x1 convolution with a sigmoid activation function to ensure the output is a probability.

U-net is a state-of-the-art segmentation model that has been used in many biomedical contexts [6, 29, 25]. The implementation of Buda et al. [6] used by this work also provided the parameters they acquired after training. It was trained to segment brain tumours from magnetic resonance imaging (MRI) data. These parameters are used to initialise the model, such that it can be fine-tuned using the 100+ dataset. This transfer learning approach has been shown to improve the performance of DNNs [23].

2.2.2. HANDLING LARGE IMAGES

A big challenge to overcome for the task is how to handle the large size of WSIs. They are too large to train most deep learning models on, since they need to retain a copy of the input of each layer in memory. If the initial input image can already be more than 20 GB, this becomes unfeasible. Two approaches that can remedy this are used in this work.

One way to reduce the memory requirements is to downsample the WSI before analysis and upsample it afterwards. In addition to lower memory requirements, it also lowers the amount of data that needs to be analysed, thus lowering computation times. The downside of this is that a significant amount of detail may be lost. However, for the segmentation of grey matter, this loss of detail should not be an issue. The grey matter is on average 2.5 millimetres thick [35]. Thus given the resolution of 0.274 μm per pixel, this results in the smallest parts of the grey matter still being approximately 1000 pixels wide. In conjunction with the fact that the shape of the grey matter is smooth and lacks highly detailed edges, downsampling becomes a viable option.

Another way to address this issue would be to use a patch-based approach to analyse each WSI [21, 25]. This means that the WSI is split into smaller patches that are analysed individually and then combined to create the fully segmented WSI. This lowers the memory requirements of the system at the cost of computation time. The patch size should be chosen based on the problem and the hardware available. A downside of the patch-based approach is that artefacts are likely to occur on the edges of the patches when they are stitched together after analysis. This can be mitigated by overlapping the patches. However, this adds computation time since part of the input will be analysed twice. Furthermore, it adds complexity to the combination step to manage multiple predictions for the same pixel. A different way to mitigate artefacts on edges would be to use a post-processing step [21].

The patch size is an important parameter to choose when using this approach. A high patch size lowers the number of edge artefacts in the final segmentation as there are fewer edges that can create artefacts. For training a lower patch size can be preferred. During training, multiple patches are used in each forward and backward step. This number of patches is called the batch size. Ideally, the batch size should be as high as possible. Then the model can learn from many different samples with each step, in-

creasing the probability that the parameters are updated in the right direction. A lower patch size frees up memory such that a larger batch size may be used. However, the patch size should not be too low. If it is, the patch might not contain enough information to accurately segment the grey matter.

This work uses both approaches to strike a balance between segmentation detail, memory requirements and computation time. Downsampling until the WSIs are small enough to be analysed in one pass would likely not leave enough detail to segment correctly. Furthermore, it would result in a coarse segmentation that might not be good enough for the plaque localisation step downstream. However, downsampling is required since the computation time would be too high otherwise. Initial exploration of this resulted in an estimated training time in the order of months, given the available hardware. Thus each WSI is downsampled 16 times before analysis. This is similar to Lai et al. [21] as the resulting segmentation map is 16 times smaller than the input as well.

For the patch-based approach, this work opts for non-overlapping patches and no post-processing step. To help train and limit the amount of artefacts different patch sizes are used for training and for creating the full segmentation. This can be done since the U-net model is fully convolutional and is therefore not limited to a fixed-size input. There might be a slight difference in the prediction of the model when the input size is increased. However, the reduction of edge artefacts when combining patches is expected to outweigh this negative effect. Thus, the model is trained using a patch size of 256 by 256 pixels and the full segmentation is made with a patch size of 1024 by 1024.

2.2.3. TRAINING
The model is trained using WSIs from the 100+ dataset since that set includes annotations for the grey matter. The WSIs are split into a training set, a validation set and a test set. Each contains 324, 46 and 92 WSIs respectively. As stated before in section 2.2.2, each WSI is split into patches of 256 by 256 pixels. During training, the patches are loaded in a randomised order to ensure variation within each batch and between batches. The batch size for training is set to 16, which is the limit for the available hardware. It uses cross-entropy loss as a loss function and the Adam optimiser. The learning rate is selected by first training the model for a single epoch on a wide range of learning rates between 0.1 and 0.00001. Based on the learning curves of those models, a selection is made that will be trained for a further 4 epochs.

2.2.4. EVALUATION
To evaluate the segmentation models, each goes through all WSIs in the test set and constructs a full segmentation of it. As stated in section 2.2.2, this is done in patches of 1024 by 1024. It is compared to the ground truth using two metrics: intersection over union (IoU) and the Dice coefficient (DC).

The IoU measures the ratio between the intersection and the union of the model prediction (P) and the ground truth label (L). The DC is similar but is the ratio between two times the intersection and the sum of the areas of the prediction and ground truth. Both result in a score between 0 and 1, where a 0 means the prediction is completely wrong and a 1 means the prediction is perfect. They are correlated, but the difference is that the IoU punishes bad instances more harshly than the DC. When averaged over

the test set, the IoU represents a worst-case performance and the DC an average-case performance.

$$IoU = \frac{|P \cap L|}{|P \cup L|} \tag{2.1}$$

$$DC = \frac{2|P \cap L|}{|P| + |L|} \tag{2.2}$$

2.3. LOCALISATION

The goal of the localisation algorithm is to locate individual Aβ plaques in each WSI. Using the result from the segmentation algorithm, the grey matter is scanned for plaques. Plaques are found by masking them and creating a bounding box using that mask. The algorithm that creates the masks is not DNN-based, but based on more traditional image processing techniques.

2.3.1. PLAQUE DETECTION

The Aβ plaques are stained using 3,3'-Diaminobenzidine (DAB), giving the plaques a distinctive red-brown colour. The goal of the detection algorithm is to find a mask for each individual plaque in the image. This mask can be used to create a bounding box for the plaque which can then be used for a downstream task. The plaque detection algorithm has several parameters for which the final values will be experimentally determined.

The first step in this process is to create a greyscale image in which the plaques should be distinct. This can be done using the colour deconvolution designed by Ruifrok and Johnston [30] and implemented by scikit-image[34]. This is an algorithm that transforms an image from the red-green-blue (RGB) colour space into the hematoxylin-eosin-DAB (HED) colour space. Each channel in the HED colour space represents a dye that is commonly used in Immunohistochemistry. The detection algorithm only needs to find DAB staining, therefore, that channel is isolated (figure 2.4b). The DAB signal is low, but still usable for the following steps.

In the second step, a threshold is applied to the DAB channel to get a binary image. This is found using Otsu's thresholding algorithm [26] as implemented by openCV [17]. This algorithm calculates the threshold by iterating over possible threshold values and calculating the weighted variances on each side of that threshold. The value that results in the highest variance between the two sides is selected as the optimal threshold. In short, it finds a threshold such that the distinction between foreground and background is most significant. The threshold can therefore be different for each input image, which is necessary to handle different staining intensities in the dataset. However, there are also cases that have no or very little Aβ pathology. In that case, the Otsu threshold will be (close to) zero and it might include a huge area in the positive class, which is incorrect. To remedy this, a minimum threshold is added as a parameter of the algorithm. When the Otsu threshold is lower than the minimum threshold, the threshold is set to the minimum threshold. The result of the thresholding step can be found in figure 2.4c The third step aims to smooth the output of the previous step. Figure 2.4c shows that there are many small holes and loose bits in and around the main bodies. This is smoothed using
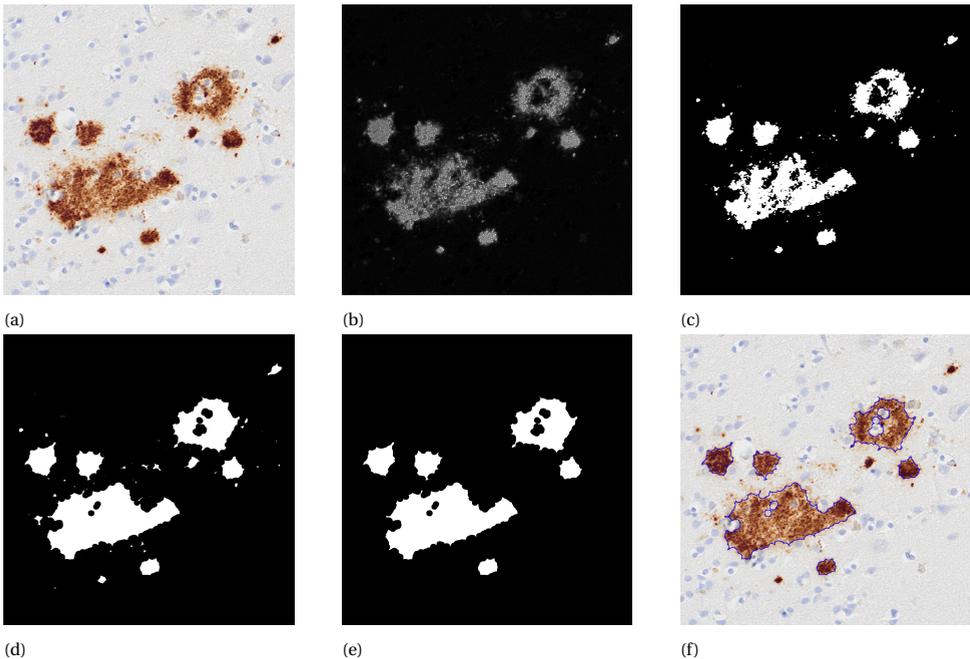
Figure 2.4: A step-by-step visual example of the plaque localisation algorithm. (a) An example input image. The size of this image is 1024 by 1024 pixels or 280 by 280 microns. (b) The DAB channel of the colour deconvolution. The contrast in this image is enhanced because the original signal is too weak to see well. (c) The binary image made using the threshold found by the Otsu algorithm (d) The binary image after closing with kernel size 21. (e) The binary image after removing all detections smaller than 10 microns in diameter. This removed 43 out of 49 initial detections. (f) The original image with the plaque boundaries drawn in blue.

a single morphological closing step using a round kernel. The size of the kernel is added as a parameter to the algorithm. A small kernel might not smooth the detections enough or leave too many loose bits. Although, if the kernel is too large it might glue plaques together that should be separate. Figure 2.4d shows the effect of the closing step.

The fourth and final step is to find all the connected components in the binary image and filter out the ones that are smaller than some minimum size, which is added as a parameter. The connected components are found using openCV [17]. The area of each component is compared to the area of a circle with a diameter equal to the minimum size. If it is smaller than the minimum size, the component is discarded. The effect of this can be seen in figures 2.4e and 2.4f. The smaller detections are removed because they are not considered pathologically interesting enough. The plaques are collected by saving the bounding box and the plaque mask. This information can be used for any downstream task.

One could argue that the closing step should be performed after the smallest objects are removed. That might result in fewer false positive detections and smaller plaques. However, it might also remove small bits that are accidentally separated from the main body of the plaque. The work chooses to do the closing step before the removal of small

detections because it is preferable to have a few more false positives than to miss (parts of) plaques.

2.3.2. PATCH-BASED ANALYSIS

The localisation algorithm has a similar problem as the segmentation algorithm: A WSI is too large to be analysed in one pass, due to the large memory requirements. Again, this can be solved by downsampling or using a patch-based approach. Since the plaques can be very small, it is not possible to downsample the input image. That would lead to too much loss of detail. A patch-based method is therefore employed. The WSI is split into non-overlapping patches of 4096 by 4096 pixels before analysis. First, the patch is compared to the grey matter masks in the dataset. If there is no grey matter in the patch, it is not analysed because Aβ pathology is only found in the grey matter and it saves computing time. If there is grey matter in the patch, the plaque detection algorithm is run.

After all patches are processed, it is time for the final step. When creating the patches, it might occur that a plaque is split between two or more patches. A plaque merging algorithm is implemented to fix this issue. The probability that the bounding boxes around two plaques exactly border each other is assumed to be extremely low. Therefore, the merging algorithm checks whether any plaques have bordering bounding boxes. When a pair is found, they are merged, which creates a new plaque mask and a new bounding box.

2.3.3. EVALUATION

To find the correct parameters for the localisation algorithm a short exploration is done to find good candidates. Several runs were done on images in the plaque detection dataset (section 2.1.3) and the results were visually inspected. This exploration resulted in the following parameter choices.

The first is the minimum threshold, for which the values of 0, 0.03, 0.04 and 0.05 are chosen. The values in the DAB channel can range between 0 and 1. However, most values are very low, thus a low minimum threshold is appropriate.

The kernel sizes to test are 1, 11, 21 and 31. A kernel size of 1 is tested to see whether the step is necessary at all. The other values represent a circular kernel with that diameter. The exploration showed that a kernel size beyond 31 would agglomerate too many plaques, thus values from 1 to 31 were included.

The minimum plaque size is set to 5, 10 and 15 microns. These are tested to get an idea of how many plaques are removed by each setting. The smallest plaques that are of interest are usually around 10 microns in diameter. Smaller plaques do appear, but they have not been thoroughly researched [33]. The 5 and 15 micron settings are added to this experiment to verify whether this holds true for this algorithm.

The localisation algorithm is evaluated by running it on the plaque detection dataset and comparing the detections with the ground truth in that set. Since it is an object detection algorithm, evaluation metrics are borrowed from this type of problem. In object detection, a detection is correct when the IoU of the detection and the ground truth label is higher than some threshold, e.g. 0.5 [22, 11]. However, the ground truth labels in the plaque detection dataset are often larger than the actual plaque (figure 2.5a). In this sit-

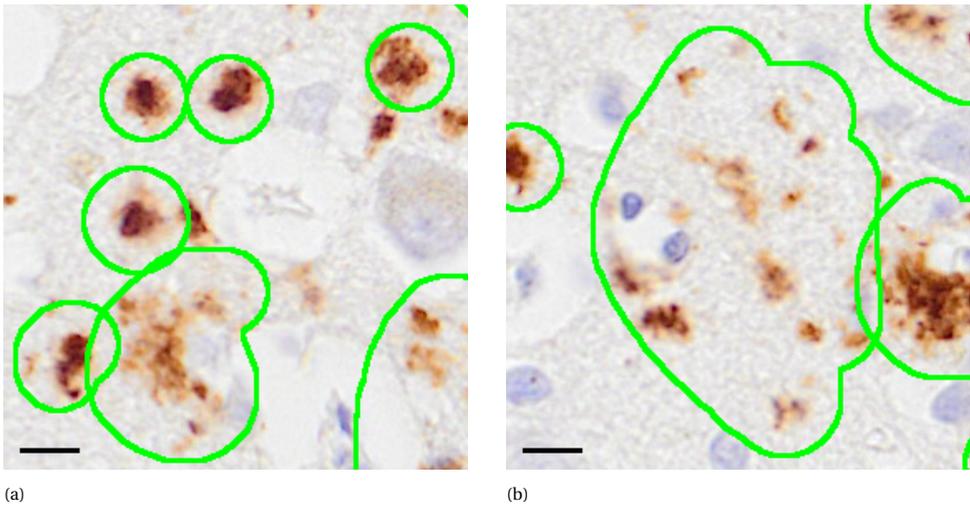(a)   (b)

Figure 2.5: Two examples of ground truth labels that show the need for using IoD instead of IoU as a criterion. (a) shows several plaques where the annotation is larger than the plaque itself. (b) shows an example where multiple small plaques are annotated as a single plaque. Bar = 10 µm in both images

uation, a plaque that should be counted as correct can easily be considered incorrect as the area of the plaque is not large enough to make the threshold of an IoU of 0.5. Therefore, this work proposes a different criterion: Intersection over detection (IoD). This is similar to the IoU only that it is more lenient towards small detections in large labels. With IoD, a detection that is fully surrounded by a label will still be counted as correct.

$$IoU = \frac{|D \cap L|}{|D \cup L|} \tag{2.3}$$

$$IoD = \frac{|D \cap L|}{|D|} \tag{2.4}$$

Another change to the traditional correctness criterion is that multiple detections per ground truth label are allowed. There are instances in the dataset that have multiple small plaques in the same label (figure 2.5b). Furthermore, the borders between plaques are not always well-defined. Different pathologists may have different ideas on whether a plaque may actually be two plaques and vice versa. Therefore, if there are multiple detections for a single label, all may be considered correct.

Using the IoD the precision and recall are calculated for each of the sets of parameters. Precision is the number of correct detections divided by the total number of detections. High precision indicates that many detections by the algorithm are good correct detections. Recall is the number of correct detections divided by the number of plaques in the ground truth. High recall indicates that the algorithm does not miss many detections that it should have picked up. These two measures give an indication of the accuracy of the detection algorithm.

2.4. CATEGORISATION

The final step of the process is the categorisation. Here the goal is to extract information from individual plaques. That information is used to see whether there are big differences in plaques between AD cases and 100+ cases, and whether any correlations exist between plaque physiology and cognition in centenarians. To do this, features are extracted from the plaques using both traditional and deep learning-based methods. Then a clustering algorithm is employed to create a label for each plaque.

2.4.1. FEATURE EXTRACTION

Each detected plaque has a mask and a bounding box. The mask contains information about the shape and size of the plaque. The bounding box has the RGB colour values for the area where the plaque is located.

Two features are extracted from the mask: the area and the roundness. The area is calculated by counting the number of pixels in the mask and multiplying that with the area of a pixel ($\sim 0.075\mu$m^2). The size of a plaque might have an impact on cognition since a larger part of the brain is affected by the plaque. However, it might also be the case that a large number of smaller plaques lead to worse cognition. This is why this feature is picked for analysis. The area can also be used to calculate the Aβ load of a WSI by summing the areas of all plaques in a WSI and dividing that by the area of the grey matter. This can be compared with the Aβ load calculations of Hernández [15] to see if the results are similar.

The roundness is calculated as the ratio between the diameters of the maximum inscribed circle and the minimum circumscribed circle of the mask [16]. This results in a metric between zero and one. If the roundness is one, it means the mask is a perfect circle. If the roundness nears zero, the mask becomes less and less round. Figure 2.6 shows some examples of the calculation of roundness. This shape feature might be able to differentiate between several types of plaques since compact and dense plaques are much rounder than diffuse plaques or subpial aggregations. The bounding box is used to create a different kind of feature. It is used as the input for a pre-trained Alexnet [20]. Alexnet is an early convolutional neural network architecture designed for image classification. It consists of five convolutional layers followed by 3 fully connected layers. The specific version used in this work is implemented by PyTorch [27] and trained to reproduce the results of Krizhevsky [19]. The network is trained on Imagenet1K [8]. This is a dataset with images of 1000 different classes, ranging from goldfish to beer glass to pipe organ. To predict the class to which an image belongs, Alexnet returns a vector of length 1000. Each entry in the vector corresponds to the probability that the image is of a specific class.

Imagenet does not contain any images of Aβ plaques, therefore the output of the Alexnet should be quite noisy as it does not know how to classify the image. However, this vector might still contain useful information about the plaque. The output vector is used as a latent representation of the plaque image, which can be used in downstream tasks.

Before the bounding boxes are passed through the Alexnet, they are resized to be 224 by 224 pixels. This is done because Alexnet is not fully convolutional and has a fixed-size input. This resizing loses some information about the size and shape of the plaque, as
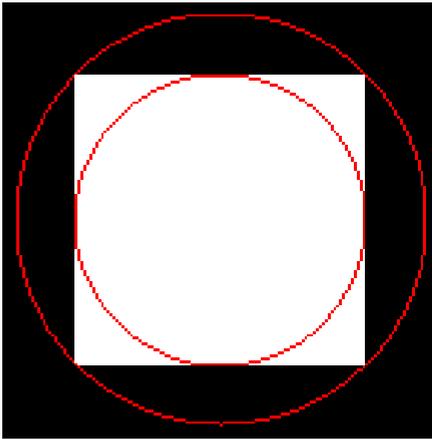
(a) Inner diameter: 100 pixels
Outer diameter: $100\sqrt{2} \approx 141$ pixels
Roundness: $\frac{100}{141} \approx 0.71$
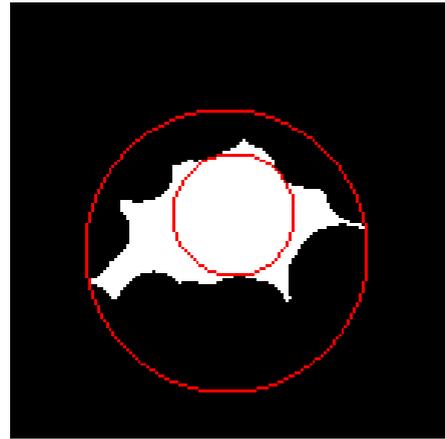
(b) Inner diameter: 42 pixels
Outer diameter: 97 pixels
Roundness: $\frac{42}{97} \approx 0.43$

Figure 2.6: Two examples of shapes with the maximum inscribed circle and minimum circumscribed circle in red. (a) shows a square and (b) an example of a plaque mask.

plaques may be distorted or downsampled. However, after inspecting the WSIs, it looks like most plaques are smaller than 224 by 224 pixels. Thus, losing information due to downsampling should not be a big issue.

2.4.2. CLUSTERING

The Alex features by themselves are not meaningful to humans. Therefore a clustering algorithm is applied to cluster the latent representations of all the detected plaques. The goal is that the created clusters resemble the different types of plaques that are known to exist in the literature. If a good clustering can be found, a profile of the plaque types in a WSI can be made. This information could shine a light on whether there are differences in plaque types between 100+ cases and AD cases. Furthermore, the different plaque types can be correlated with the cognition scores in the 100+ dataset, showing the possible impact different plaque types have on cognition.

K-means clustering is picked as the clustering algorithm. This algorithm creates an initial set of k means in the feature space. Then all samples are assigned to the mean they are closest to, resulting in k sets of samples. For each set a new mean is calculated and the samples are re-assigned to the mean they are closest to. This is repeated until convergence. This results in k clusters that have close to equal variance.

The K-means clustering algorithm scales well in the number of samples due to its iterative nature and the possibility for parallel processing of samples [4]. This is necessary since the number of plaques in the 100+ and AD datasets could be in the millions. Furthermore, the algorithm has few parameters to set, only the number of clusters k, making it easy to use.

The implementation used in this work is provided by Scikit-learn [28]. The initial means are selected using the K-means++ algorithm. This selects the initial means to be

mostly well-separated, often leading to better results than using a random initialisation. 5, 10 and 20 are the number of clusters that are chosen to be tested. The minimum number of different plaque types should not be lower than 5 as that would not be enough to include the possible types of plaques. K-means clustering creates clusters that have a similar number of samples in them. Some plaque types, such as CAA plaques, occur orders of magnitude less frequently than dense plaques. Therefore, a low number of clusters might not be fine-grained enough to find these rare plaques. Thus 10 and 20 are tested as well.

The results of the clustering are evaluated by trained pathologists. They shall look at a sample of plaques from each cluster and assess whether the grouping looks valid.

3

RESULTS

3.1. SEGMENTATION RESULTS

The goal of the segmentation step is to label the grey matter in a WSI. For this task, a U-net was trained with various learning rates on patches from WSIs to predict which parts contained grey matter. The best-performing model is chosen to fully segment WSIs. In order to do that, the WSI is first split into patches. Each patch is segmented individually by the U-net and they are combined to create the final result.

3.1.1. LEARNING RATE

The results of the initial learning rate experiment can be found in figure 3.1. This graph clearly shows two groups of learning rates: 4 that perform better and 4 that perform worse. It is not possible to select 1 best learning rate based on this experiment alone, thus the next experiments are performed using learning rates equal to 0.001, 0.0005, 0.0001 and 0.00005.

Table 3.1 shows the evaluation of the trained models after 1 epoch and 5 epochs of training. According to this table, the model with a learning rate of 0.0001 after 5 epochs seems to perform best. Both the mean IoU and the mean DC are highest and the standard deviation the lowest, indicating that the results are the best and the most consistent. Therefore, this version is picked as the final model. All the model predictions shown in figures 3.2-3.5 are made using this model.

3.1.2. FULL WSI SEGMENTATION

The overall results of the segmentation algorithm are good, as the scores in table 3.1 suggest. Figure 3.2 shows that the model is capable of producing an accurate segmentation. The biggest difference is visible on the bottom where the model has labelled a piece of the meninges as grey matter.

Figure 3.3 shows an example that appears to be segmented badly. Grey matter is detected on the bottom part of the slide, but it has holes that are not visible in the ground truth. However, there is an increased amount of nuclei in the locations of the holes,
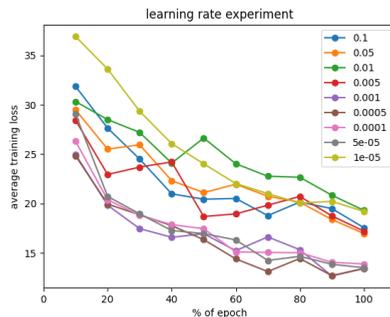
3

Figure 3.1: The training loss for the first epoch. Each point represents the average loss over the previous 10% of the epoch, which is approximately 46 WSIs.

Table 3.1: IoU and DC scores for the selected learning rates after training for 1 epoch and 5 epochs. Each entry shows the average score over the test set and the corresponding standard deviation. The best-performing model is shown in bold.

Learning rate	1 epoch		5 epochs	
	IoU	DC	IoU	DC
0.001	0.2125 ± 0.1988	0.3096 ± 0.2519	0.7193 ± 0.2063	0.8160 ± 0.1744
0.0005	0.1230 ± 0.2076	0.1709 ± 0.2621	0.7437 ± 0.2310	0.8267 ± 0.2010
0.0001	0.8801 ± 0.0712	0.9346 ± 0.0441	$\mathbf{0.8962 \pm 0.0706}$	$\mathbf{0.9436 \pm 0.0441}$
0.00005	0.7637 ± 0.1824	0.8514 ± 0.1424	0.8883 ± 0.0737	0.9390 ± 0.0469

(a) WSI   (b) Ground truth   (c) Model prediction
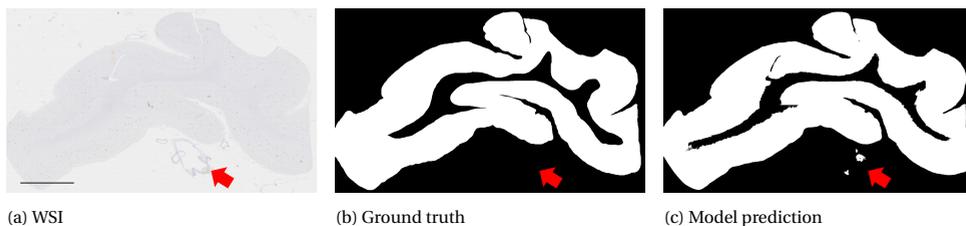
Figure 3.2: A WSI of tissue from the temporal pole, the ground truth for the grey matter and the segmentation by the model. It shows a generally accurate segmentation, except for a piece of meninges that is accidentally labelled as grey matter (indicated by the red arrow). Bar = 5mm for all three images.

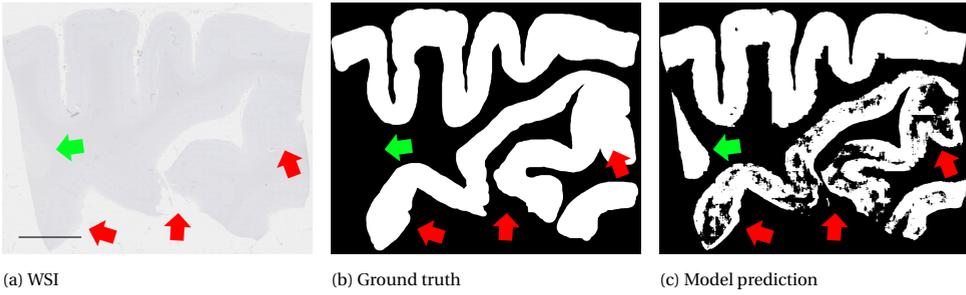(a) WSI   (b) Ground truth   (c) Model prediction

Figure 3.3: A WSI of tissue from the occipital pole, the ground truth for the grey matter and the segmentation by the model. The red arrows indicate an area where white matter appears in the grey matter. This is picked up by the model, but omitted by the annotator. The green arrow indicates an area with grey matter that was correctly detected by the model, but not annotated. Bar = 5mm for all three images.

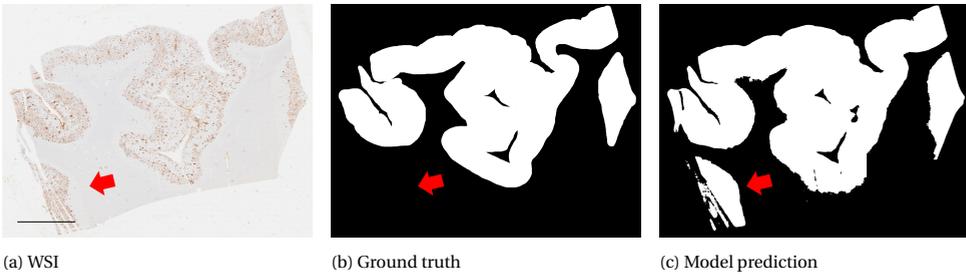(a) WSI   (b) Ground truth   (c) Model prediction

Figure 3.4: A WSI of tissue from the parietal lobe, the ground truth for the grey matter and the segmentation by the model. This WSI is heavily stained and is segmented well despite that. The red arrow indicates an area with grey matter that was correctly detected by the model, but not annotated. Bar = 5mm for all three images.

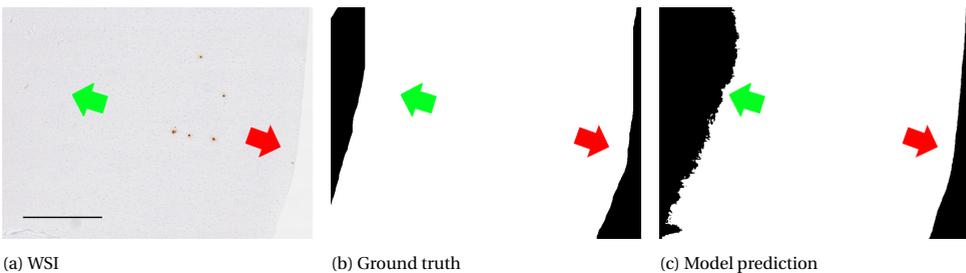(a) WSI   (b) Ground truth   (c) Model prediction

Figure 3.5: A WSI of the cortex from tissue from the temporal pole, its corresponding ground truth and prediction. The prediction has a rough inner border (green arrow) and a smooth outer border (red arrow). Bar = 1mm for all three images.

which is characteristic of white matter. This slide was most likely improperly cut, caus-ing some white matter to appear there. While it was not noticed by the annotator, it is detected by the model. On the left side of figure 3.3, the model detects an extra piece of grey matter that is not labelled in the ground truth. However, inspection of the WSI reveals this is an error in the annotation and not a mistake by the model.

Figure 3.4 shows a heavily stained WSI. Despite that, the model still manages to make a good prediction. Moreover, it finds an extra piece of grey matter that was not anno-tated, similar to figure 3.3.

Figure 3.5 shows a part of the cortex. The border between the grey matter and the background (the outer border) is smooth in the prediction, while the border between the grey and white matter (the inner border) is rough. This might be because the outer border is much more well-defined in the input image, while the inner border is much more gradual. Even for an expert, it is hard to define the exact location of the inner border. Therefore, the prediction of the inner border is likely more realistic than the inner border in the ground truth.

Overall, the segmentation has good performance. It is able to detect the grey matter in lightly stained as well as heavily stained WSIs. Areas of grey matter that were skipped by the annotator are detected by the model (figures 3.3 and 3.4), showing its power. It is also able to detect areas of white matter in unexpected places (figure 3.3). However, due to the difficulty of defining the border between grey and white matter, it is hard to say how accurate this is exactly.

3.2. Localisation results

The goal of the localisation algorithm is to detect individual plaques in images of grey matter. This is done by applying a sequence of traditional image processing steps. The algorithm has three parameters for which a value needs to be found: the minimum threshold, the kernel size and the minimum plaque size. The settings are evaluated based on the performance of the algorithm on the plaque detection dataset (section 2.1.3) and by manual inspection of the impact of the parameters.

3.2.1. Impact of parameters

The influence of the localization algorithm's parameters must be assessed in order to select the appropriate values for them. This is done by first looking at examples, which helps to understand the effect on the result. Then it is followed by explaining the effect on the evaluation metrics in tables 3.2 and 3.3.

The effect of the minimum threshold is most evident on patches with little to no staining or very low-intensity staining. In images where the stains are dark and plentiful, the effect is not visible because the Otsu threshold is well above the minimum threshold. Figures 3.6a and 3.6b show an example of a lightly stained WSI. It is evident that a low minimum threshold results in many false detections on parts of the image that are not actually DAB stains, such as cell nuclei. However, the other extreme in figure 3.6b misses most of the lightly stained plaques.
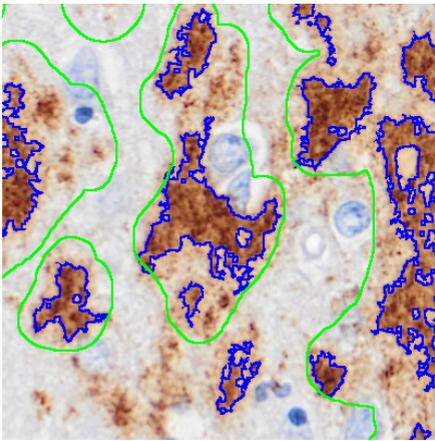
Table 3.2 shows that increasing the minimum threshold has a positive on the preci-sion if other parameters are retained. A minimum threshold of 0.05 seems the best in

3
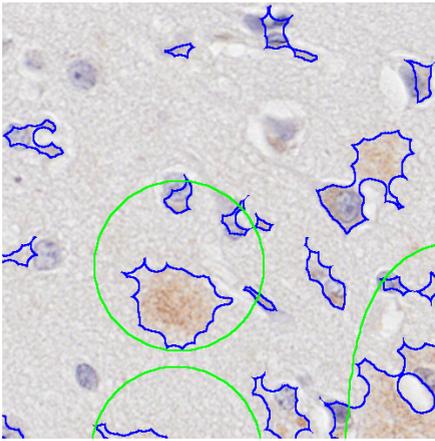
(a) Minimum threshold: 0
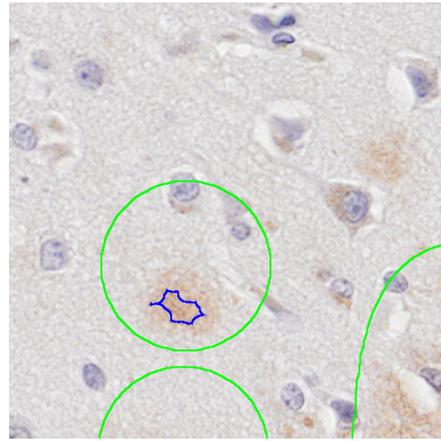
(b) Minimum threshold: 0.05

(c) Kernel size: 1
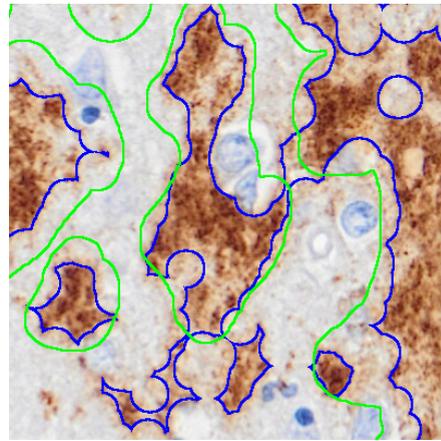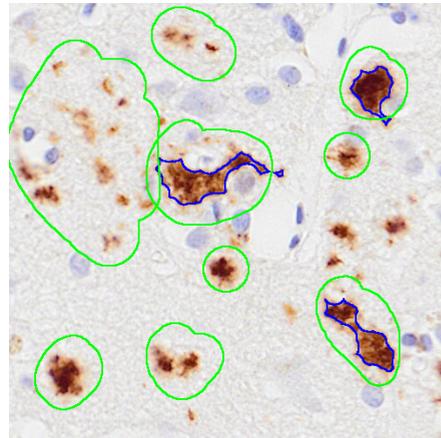
(d) Kernel size: 31
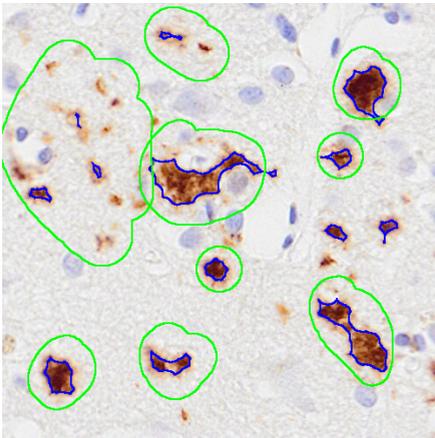
(e) Minimum plaque size: 5

(f) Minimum plaque size: 15

Figure 3.6: A comparison of the plaque detection algorithm for various parameters. The baseline settings for each depiction are a minimum threshold of 0.04, kernel size of 21 and minimum plaque size of 10. Deviations are mentioned in the caption. Detections and labels are shown in blue and green respectively.

Table 3.2: Precision of the plaque detection algorithm on the plaque detection dataset

Precision		Kernel size			
Minimum plaque size	Minimum threshold	1	11	21	31
5	0	0.106	0.380	0.394	0.401
	0.03	0.444	0.481	0.506	0.517
	0.04	0.469	0.522	0.566	0.579
	0.05	0.478	0.545	0.584	0.606
10	0	0.307	0.638	0.607	0.576
	0.03	0.694	0.734	0.758	0.758
	0.04	0.711	0.753	0.783	0.790
	0.05	0.720	0.762	0.792	0.799
15	0	0.510	0.774	0.745	0.695
	0.03	0.791	0.818	0.837	0.842
	0.04	0.801	0.819	0.851	0.855
	0.05	0.815	0.829	**0.865**	0.861

general. Table 3.3 shows that a lower minimum threshold is generally beneficial in terms of recall. This is to be expected since a lower threshold allows the algorithm to select a larger area and miss fewer of the labels.

Increasing the kernel size has two main effects on the result: smoothing and agglomeration. In figure 3.6c one can see that the detections are rough and not a good representation of the plaque per se. Furthermore, it detects some plaques as separate while they should be a single plaque. Figure 3.6d shows that the borders of the detections look much better. However, it also shows that it has agglomerated some plaques a little too eagerly. The middle plaque has been connected to the one on the right, which is clearly incorrect.

A larger kernel size has a mostly positive effect on the precision, as can be seen in table 3.2. The effect of the kernel size is less noticeable in the recall scores in table 3.3. The table clearly shows that the closing step is necessary since a kernel size of one is almost always the worst in both metrics. The increased agglomeration that comes with a larger kernel does not seem to be represented in these scores. If agglomeration was a big problem both the precision and recall should get lower when the kernel size increases. However, this is only seen in the recall scores for minimum plaque sizes 5 and 10. Here the score drops off a little for kernel size 31.

Figures 3.6e and 3.6f show the difference between the settings for the minimum plaque size. smaller plaques such as the ones on the bottom are no longer counted if the minimum size is higher. For this parameter, a balance is probably best.

Table 3.2 shows that the precision increases greatly when the plaque size is increased. This can be explained by the fact that an increased minimum size will remove many small detections that are not always annotated. This decreases the number of false detections, therefore increasing the precision. For the recall, the exact opposite is true, which can be seen in table3.3. Increasing the minimum size removes small plaques that are annotated. That increases the number of false negatives, thus decreasing the recall.

Table 3.3: Recall of the plaque detection algorithm on the plaque detection dataset.

Recall		Kernel size			
Minimum plaque size	Minimum threshold	1	11	21	31
	0	0.830	**0.856**	0.851	0.819
5	0.03	0.821	0.850	0.848	0.821
	0.04	0.788	0.820	0.823	0.798
	0.05	0.754	0.787	0.789	0.768
	0	0.647	0.684	0.691	0.680
10	0.03	0.636	0.677	0.687	0.681
	0.04	0.597	0.638	0.657	0.648
	0.05	0.570	0.610	0.625	0.618
	0	0.443	0.491	0.512	0.520
15	0.03	0.429	0.480	0.507	0.520
	0.04	0.400	0.444	0.479	0.487
	0.05	0.384	0.423	0.447	0.455

3.2.2. FINAL PARAMETER CHOICES

The final choices for the parameters are based on the discussion in the previous section and on visual inspections by an expert pathologist. The scores in tables 3.2 and 3.3 do not show the full picture, since the labels and metrics used are flawed. They serve as a general indication of the direction. Furthermore, the parameter values are tailored to this specific dataset. A dataset with different levels of staining may require different parameter settings to get the best results.

The easiest parameter to pick is the minimum size, which should be set to 10. This strikes a balance between the precision and recall scores since the precision is very bad on 5 and the recall is very bad on 15. This also conforms to the expectation that plaques of interest are usually at least 10 microns in diameter.

The minimum threshold for this dataset should be set to 0.04. This choice is a balance between a good recall and a good precision score, since precision is mostly best on 0.05 and recall on 0. 0.04 was chosen over 0.03 because the latter resulted in more false positives when the patches were mostly empty (figure 3.7). This type of patch is underrepresented in the plaque detection dataset compared to the 100+ dataset and, to a lesser extent, the AD dataset. Especially the 100+ set has many areas that have little to no pathology at all. Therefore, a minimum threshold of 0.04 is preferred.

The recommendation for the kernel size is 21. The precision is higher for kernel size 31, but the recall is better for 21. This, combined with the possibility of creating large agglomerations for kernel size 31 (figure 3.6d), justifies picking the lower kernel size.

3.2.3. PLAQUE DETECTIONS IN FULL WSIS

With the parameters described in section 3.2.2, the plaque detections algorithm returned 4,746,533 detections. Table 3.4 shows several statistics about the number of plaques in the WSIs and figure 3.8 shows how the plaques are distributed across the analysed WSIs.

These results show that the AD cases contain on average almost four times as many plaques than the 100+ cases. However, 41% of slides in the 100+ set contain fewer than
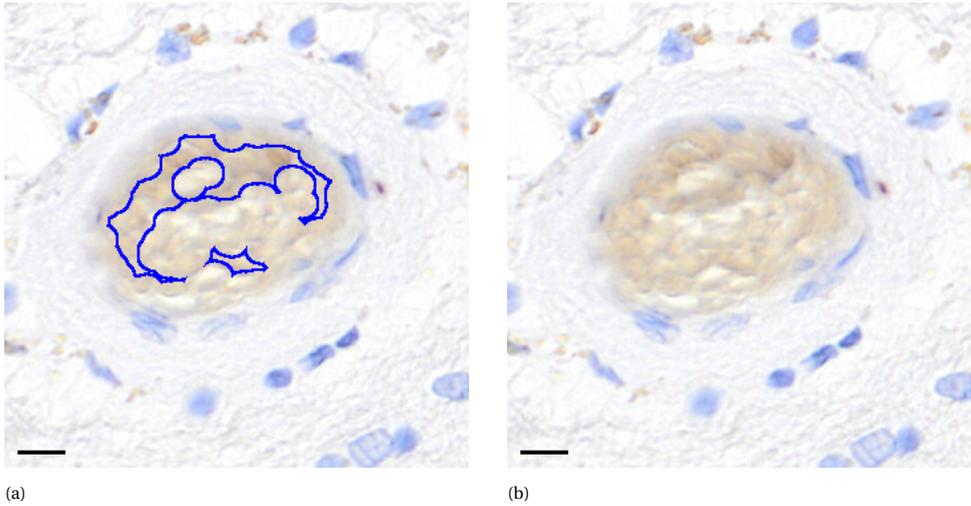
3

(a)   (b)

Figure 3.7: An example cropped from one of the images in the plaque detection dataset with no Aβ present. Both images show the result of the localisation algorithm, (a) uses a minimum threshold of 0.03 and (b) uses a minimum threshold of 0.04. It shows that 0.04 is less likely to return false positives. Bar = 10 μm in both images.

(a)   (b)
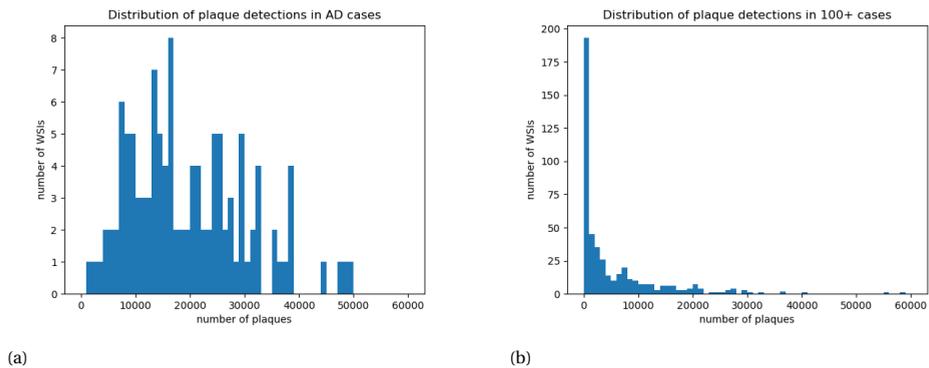
Figure 3.8: The distribution of the number of plaques found in each WSI. The set is split between AD cases on the left and 100+ cases on the right.

Table 3.4: Statistics on the number of plaques found in the WSIs. The median, mean and standard deviations are calculated over the WSIs in each dataset.

	AD	100+
Total	2278890	2467643
Median	16950	1754.5
Mean	19645.6	5341.2
Standard deviation	10656.7	8026.1

1000 plaques, meaning a couple of outliers skew the average. The median number of plaques is therefore a better statistic to use. That tells us that the median AD case contains roughly 10 times as many plaques as the median 100+ case. These counts show that AD cases have more pathology than 100+ cases.

3.3. Categorisation results

The goal of the categorisation is to collect information on individual plaques and use that information to differentiate between different types of plaques. This is done by extracting a shape and size feature from the masks of detected plaques and comparing them between slides. In addition, a feature vector is extracted from each detected plaque using a CNN. The collection of vectors is clustered to create categories of plaques.

3.3.1. Aβ load comparison.

Figure 3.10 shows a comparison between the Aβ loads as calculated by this work and the unpublished data by Hernández [15]. The plots alone suggest a correlation between the two, which is confirmed by the correlation coefficients and corresponding p-values. From the plots, it can be noted that the load calculated by this work is lower than the unpublished work. A direct comparison of the results of the detection algorithm by Hernández and the localisation algorithm in this work reveals why. The detection of Hernández is done at a lower resolution and includes more of the surrounding area of the plaques (figure 3.9). The localisation algorithm in this work creates a more detailed mask that includes less extra space.
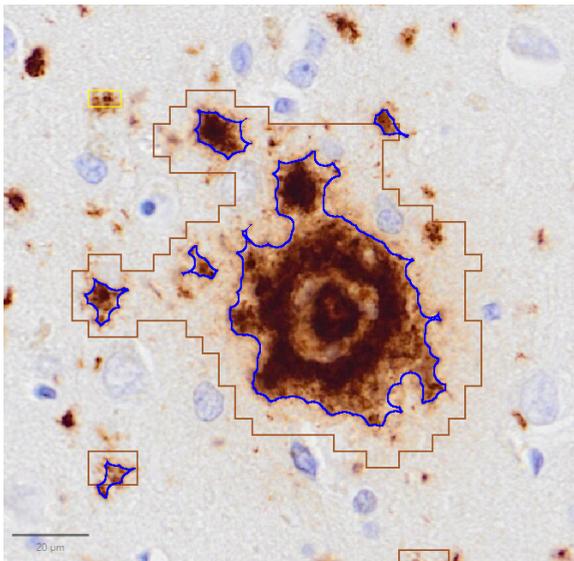

Figure 3.9: A comparison of plaque detection by Hernández [15] in brown and the localisation algorithm by this work in blue. The total areas of the brown and blue detections in this image are approximately 6100μm^2 and 2400μm^2 respectively. Bar = 20 μm

3

(a) Correlation coefficient:0.949
p-value: ≪ 0.001

(b) Correlation coefficient:0.964
p-value: ≪ 0.001

(c) Correlation coefficient:0.784
p-value: ≪ 0.001

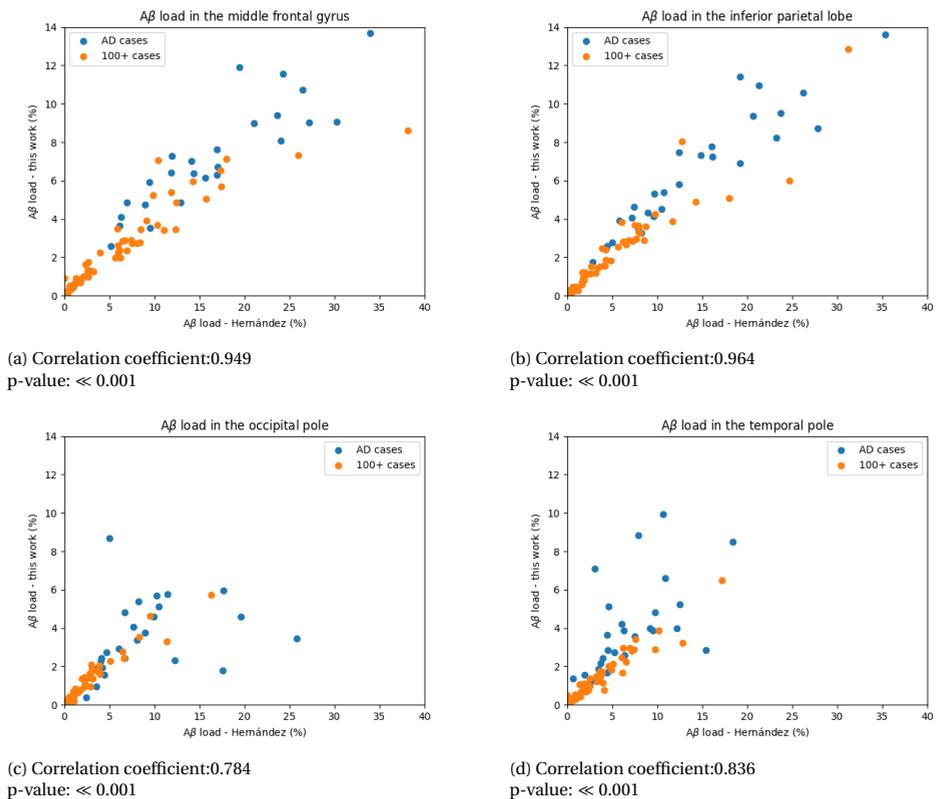(d) Correlation coefficient:0.836
p-value: ≪ 0.001

Figure 3.10: A comparison of the Aβ load in four brain regions analysed by this work and the unpublished work. The values on the axes represent the percentage of the grey matter area that is covered by Aβ plaques. Blue points are WSIs from AD cases and orange points are WSIs from 100+ cases. The Pearson's correlation coefficient and its corresponding p-value are displayed below each plot.
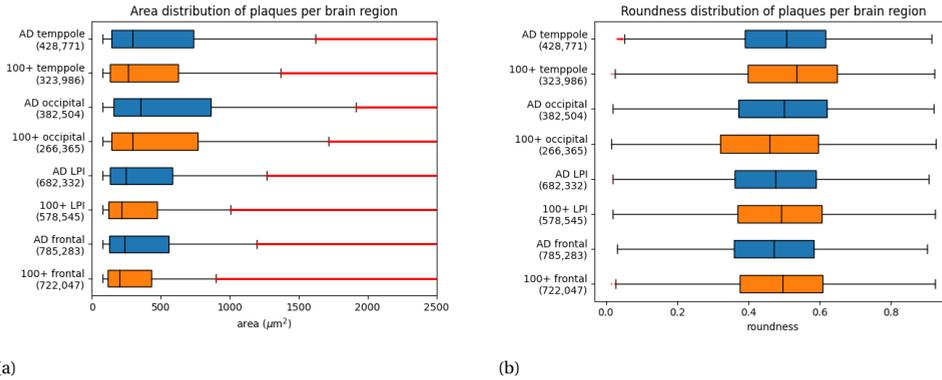
(a)   (b)

Figure 3.11: Distributions of the plaque area (a) and the plaque roundness (b). For this the plaques were divided based on the brain region they were found in and whether they came from an AD case (blue) or 100+ case (orange). The number below each label on the y-axis is the number of plaques found in each set. The red dots in the boxplots are outliers. To keep plot (a) readable, the x-axis is limited to 2500 μm^2. This includes 96.2% of the detected plaques. However, plaques can get larger, as the largest detected plaque has an area of 1.02mm^2.

3.3.2. AREA AND ROUNDNESS

Figure 3.11a shows how the area of Aβ plaques are distributed per region and case. The plaques in AD cases seem generally larger than the ones in 100+ cases for each brain region. This suspicion is confirmed by a generalized Wilcoxon test. This tests the null hypothesis that the probability that a random sample from the AD set is larger than a random sample from the 100+ set equals 0.5. For each region, the p-value was lower than 0.001, which indicates the results are significant. To get an impression of the sizes of plaques, figure 3.12 shows examples of plaques of various sizes.

The distribution of the roundness of plaques can be seen in figure 3.11b. There do not seem to be big differences in the distribution, especially when comparing the same region in AD cases and 100+ cases. However, the small differences that exist are significant when tested using a generalized Wilcoxon test (all p-values below 0.001). The average plaque roundness in AD cases is 0.01-0.02 smaller than in 100+ cases, except for the occipital pole, where the average is 0.03 larger in AD cases. Figure 3.13 shows several examples of the roundness of plaques.

The next step is to see whether there are correlations between cognition scores and plaque area and roundness. To do this, the median area and roundness are taken from each brain region in each individual. A short exploration of the correlations revealed that the occipital pole showed several significant correlations which can be found in table 3.5. The digit span-backwards test and the clock drawing test show a significant negative correlation with the area and roundness of the plaques. Suggesting that individuals who perform well on those tests have smaller and less round plaques.

Table 3.5: The correlation matrix that compares the median area and median roundness in the occipital pole with the cognition scores. No corrections were made on the data. The p-values are calculated using a Student's t-test. A correlation with a p-value between 0.05 and 0.01 is marked * and a p-value between 0.01 and 0.001 is marked **.

	MMSE	Digit span-backward	Key search	Clock drawing test
Median area	-0.114	-0.232*	-0.189	-0.226*
Median roundness	-0.190	-0.325**	-0.083	-0.271*

(a) 25th percentile
Area = $128 \mu m^2$
Bar = 2 μm

(b) 50th percentile
Area = $242 \mu m^2$
Bar = 2 μm

(c) 75th percentile
Area = $573 \mu m^2$
Bar = 5 μm

(d) 99th percentile
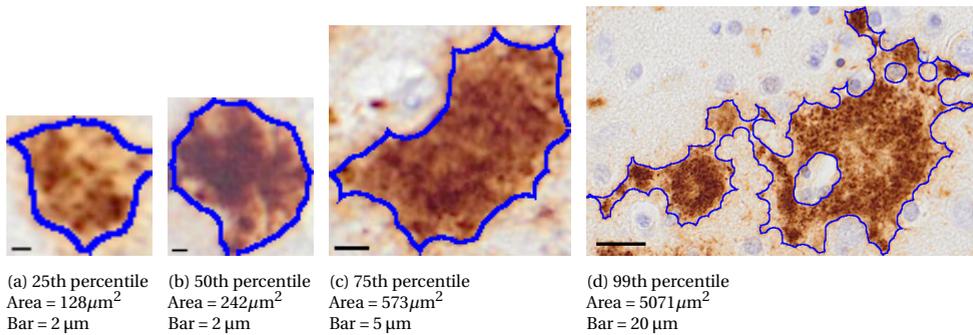Area = $5071 \mu m^2$
Bar = 20 μm

Figure 3.12: A selection of plaques to review their size. Examples are chosen by evaluating their area and selecting those at the 25th, 50th, 75th and 99th percentiles.

(a) 25th percentile
Roundness = 0.368
Bar = 5 μm

(b) 50th percentile
Roundness = 0.490
Bar = 2 μm

(c) 75th percentile
Roundness = 0.605
Bar = 5 μm
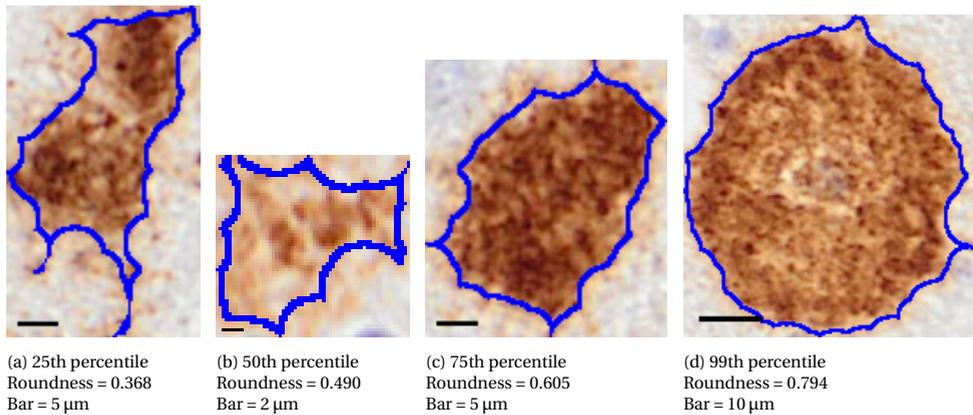
(d) 99th percentile
Roundness = 0.794
Bar = 10 μm

Figure 3.13: A selection of plaques to review their size. Examples are chosen by evaluating their roundness and selecting those at the 25th, 50th, 75th and 99th percentiles.

3.3.3. CLUSTERING RESULTS

Figure 3.14 shows the results of the clustering of the Alex features of the plaques. The results are not convincing. The goal that the clusters should resemble known plaque types is not reached. The columns represent the clusters that were found. However, no column contains plaques that strongly look alike. Therefore, the clustering results are not used for further analysis.

(a) Number of clusters: 5

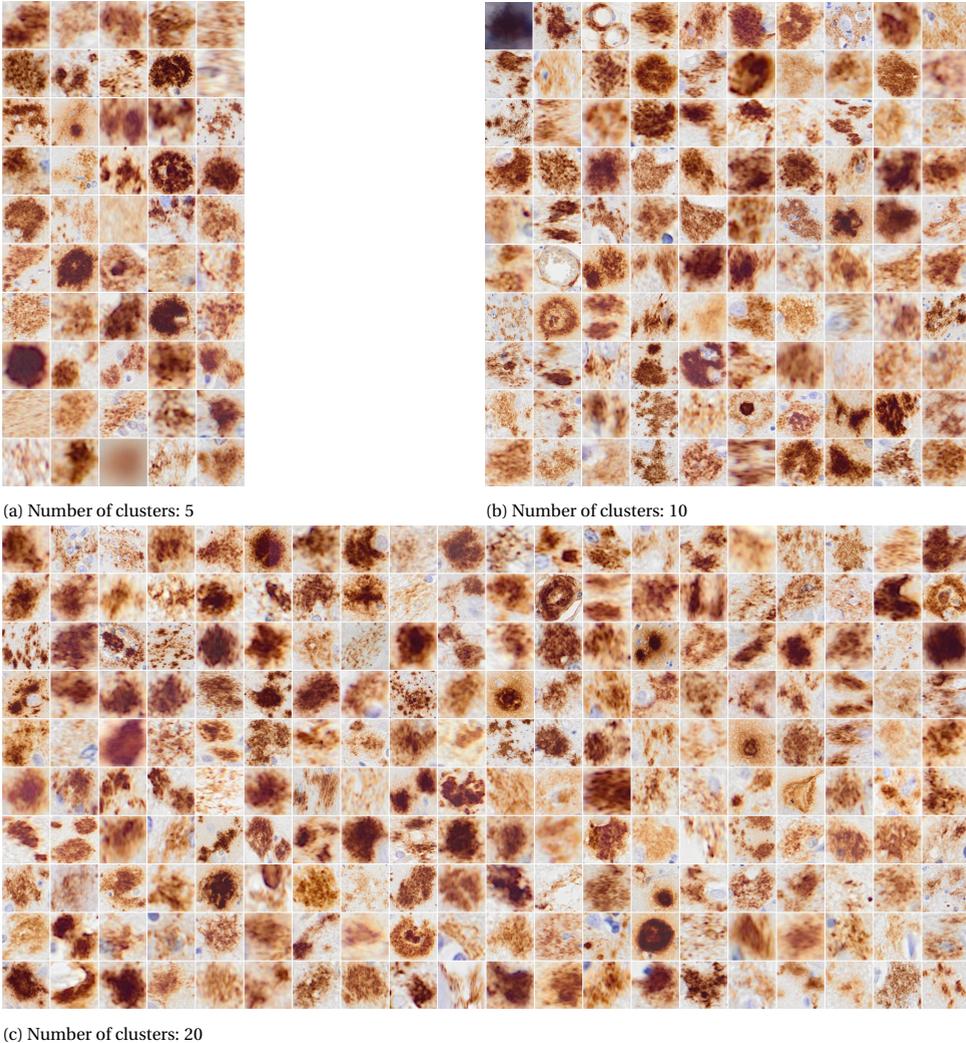(b) Number of clusters: 10

(c) Number of clusters: 20

Figure 3.14: The results of the clustering algorithm for different numbers of clusters. Each column contains ten plaques that were randomly selected from each cluster. Each plaque is reshaped to a square to resemble the input of the Alexnet.

4

DISCUSSION

4.1. SEGMENTATION

The results from section 3.1 indicate that the segmentation algorithm performs well in general. It is able to outperform human annotators in some instances (figures 3.3 and 3.4). This, however, does not mean that improvements cannot be made.

First, the training data could be improved, as is true for any deep learning-based algorithm. There are many mistakes in the annotations in the 100+ dataset. More accurate annotations could improve the performance of the model, although it is hard to say how much. The model already showed that it can perform well despite the mistakes in the annotations. There are no mistakes in the segmentation that can be easily attributed to a lack of data quality. Furthermore, improving the dataset can only be done by trained pathologists, making it very expensive to do so. Therefore, improving the dataset is unlikely to be worth the cost of doing so.

Second, a post-processing step could be added. This would mainly deal with edge artefacts and small false detections. Lai et al. [21] have shown that such a post-processing step can be beneficial. They used a traditional image processing approach not too dissimilar to parts of the localisation algorithm designed in this work. They used morphological operations to smooth artefacts and connected component detection to remove small detections. A step like this would remove false detections such as the detected meninges in figure 3.2.

Third, this algorithm (and others like it) should be more accessible to the pathology community. While this algorithm is available, not all pathologists possess the technical skills to use it. Therefore, it should be integrated with a toolbox such as Qupath [5]. Qupath supports the creation of user-made extensions packaged as java archive (JAR) files. The segmentation algorithm could be reworked to fit this format, such that it can be called directly from Qupath. Ideally, such an extension should allow more models to be added easily. This would speed up work for pathologists by giving them access to pre-trained models for segmenting different kinds of tissue, not only grey matter.

4.2. Localisation

The localisation algorithm showed that it is capable of finding Aβ plaques in various conditions with generally good results. However, there are many directions in which the localisation algorithm could be taken further.

The first would be to improve the plaque detection dataset. Both the quantity and quality of the annotations could be improved upon. The 12 patches that are annotated comprise less than 0.001% of the total dataset. Therefore, the evaluation set is not exhaustive enough to draw definitive conclusions about the performance of the algorithm. Furthermore, it would help if the annotations were more precise than they are right now. The annotation should follow the plaque edges closer and there should be just one plaque in each annotation. That would allow for a more traditional evaluation method compared to the method based on IoD (section 2.3.3), thus improving the validity of the evaluation. The problem in making these improvements to the dataset is that it will take time and resources from trained experts. The current set was annotated in 3 hours, meaning that roughly 10 plaques were annotated per minute. Making more accurate annotations is estimated to lower the annotation speed to 2-3 plaques per minute. Thus, increasing the time to recreate the set by up to five times.

Secondly, improvements can be made to the current plaque detection algorithm. The algorithm can be improved in terms of processing time. Efficiency was not the focus during the development of the algorithm, because it is a prototype and not a piece of production software. It takes between 0.5 and 1.5 hours per WSI depending on the size of the WSI and the number of plaques in it. Improvements can be made by rethinking the way in which detections are handled and stored. Furthermore, the algorithm could be improved in quality. The detections on lightly stained WSIs are not on par with the detections on darker stained WSIs. This would require a different way to do the initial thresholding step, because the lightest plaques are not distinct enough in the DAB channel. A different method to transform the RGB channels to a DAB channel can be designed such that the signal from the DAB is stronger.

A third direction would be to replace the algorithm entirely. The detections found by this algorithm could be used to create a dataset by pruning the detections by hand. The huge collection of plaques may then be used to train a DNN that might be able to outperform this more traditional algorithm.

4.3. Categorisation

The results from the categorisation are not as good as was initially hoped. The area and roundness of plaques seem to give some insight into the types of Aβ plaques in brain tissue. However, the goal to create a clustering of the plaques that resembles the plaque types described in section 1 proved infeasible with the proposed method.

The distributions of the plaque area and roundness differ between AD cases and 100+ cases. Although the results are statistically significant, it is hard to say whether the differences are relevant, because they are small. A more in-depth inspection of individual plaques in both datasets could shine more light on the cause of the small differences.

The correlations found between cognition, roundness and area of plaques in the occipital pole are interesting. It suggests that the size and shape of the plaques are related

to cognitive performance. However, the correlations are found using the median of all the plaques in a slide. This measure can be unreliable if there are only a few plaques present in a slide. A better descriptor of the distribution of area and roundness should be used to find relations with the cognitive data. A possible approach could be to select a set of measures of the distribution (mean, median, standard deviation, interquartile range, etc.) and use a regression model to find relations in the data.

However, even without a better model to find relations in the data, further exploration of it is still possible. The correlations in table 3.5 indicate that individuals who perform badly on the digit span-backwards test and clock drawing test tend to have larger and rounder plaques. Those cases could be compared to AD cases to see whether there are similarities between them.

The results of the clustering approach are disappointing, which has two main causes. The features from the Alexnet are simplistic at best. The network is not trained on images of Aβ plaques, thus the amount of information in the features is limited. A better approach could be to train an auto-encoder model on the plaques. This unsupervised deep learning method should be able to extract better representations than the pre-trained AlexNet since it actively learns from the dataset instead of only being pre-trained.

Furthermore, K-means clustering is not the best clustering algorithm for finding plaque types. First of all, K-means clustering creates clusters with roughly equal sizes. After looking at a couple of WSIs, it is very clear that plaque types are not distributed evenly. Other clustering algorithms were considered while building this algorithm, namely DB-SCAN [10] and hierarchical clustering. They are more likely to find clusters of different sizes if they exist in the data. However, they put almost all plaques in the same cluster, leaving the other clusters with only one or two plaques. If better features can be extracted from the plaque images, it could be worth it to revisit these algorithms.

Another interesting option might be to use DeepCluster [7]. This is an algorithm that combines learning features with a clustering algorithm. It first creates simple features using a convolutional neural network (CNN). These feature vectors are then clustered and the resulting cluster assignments are used as labels to train a classifier attached to the CNN, training both. The CNN is then used to generate new features, which are clustered to get new labels, which are used to train the classifier. This process is repeated until a good classifier is found. The approach can be implemented using different CNN architectures and clustering algorithms. However, the authors initially used an Alexnet and k-means clustering, with good results. Adapting the algorithm from this work to a DeepCluster approach should be feasible.

5

CONCLUSION

As the world population gets older, the prevalence of neurodegenerative diseases such as AD, increases. AD is characterised by various pathologies, one of which is Aβ plaques. To learn more about the influence of different types of Aβ plaques, this work created a system that can automatically quantify Aβ plaque types in WSIs. The system is used on a dataset from the 100+ study to see whether there are correlations between types of plaques and cognitive performance. The system is also used on a dataset of AD cases to see whether they have differences in the types of plaques present when compared with the 100+ cases.

The system has three main steps. The first step consists of an algorithm that can segment the grey matter from the WSI. This is done by applying a U-net with a patch-based approach to create a full segmentation of the WSI. In the second step, the plaques are located using a sequence of traditional image processing steps. The detections were analysed and categorised in the third step. The area and roundness of the plaques were extracted as well as an AlexNet-based feature vector. The feature vectors were clustered using k-means clustering to create plaque categories.

The segmentation algorithm shows generally good results. It is able to find almost all of the grey matter present in the WSI with few glaring errors. The localisation algorithm also performs well, finding 4.7 million plaques across the two datasets. The final categorisation step proved more difficult. The area and roundness were extracted well, but the results of the clustering of the feature vectors were not usable.

The area and roundness of the plaques are distributed differently in AD cases and 100+ cases, although the difference is small. Furthermore, the area and roundness of plaques in the occipital pole show a negative correlation with cognitive performance. These findings indicate there is information in the shape and size of Aβ plaques, opening the door for further research into this topic. More information can be found in the data since it is not yet fully explored.

The clustering approach by this work failed, but it should not be given up. The AlexNet features are basic because the model was not trained with Aβ plaques in mind. Another limitation is the use of the k-means clustering algorithm. It was unable to find

a meaningful structure in the AlexNet feature. Future research should find a more so-phisticated method to extract a vector representation of the plaques as well as a better method for assigning them to categories.

In conclusion, this work created a system that can accurately locate Aβ plaques in WSIs of brain tissue. The analysis of the plaques indicated there may be a relationship between the type of Aβ plaques in the brain and the cognitive performance of an indi-vidual. However, further research is necessary to unravel the nature of this relationship.

5

BIBLIOGRAPHY

[1] Feb. 2023. URL: https://100plus.nl/.

[2] Alzheimer's Association. "2019 Alzheimer's disease facts and figures". In: *Alzheimer's & dementia* 15.3 (2019), pp. 321–387.

[3] Alzheimer's Association. "2019 Alzheimer's disease facts and figures". In: *Alzheimer's & dementia* 15.3 (2019), pp. 321–387.

[4] Bahman Bahmani et al. "Scalable k-means++". In: *arXiv preprint arXiv:1203.6402* (2012).

[5] Peter Bankhead et al. "QuPath: Open source software for digital pathology image analysis". In: *Scientific Reports* 7.1 (Dec. 2017). DOI: 10.1038/s41598-017-17204-5. URL: https://doi.org/10.1038/s41598-017-17204-5.

[6] Mateusz Buda, Ashirbani Saha, and Maciej A Mazurowski. "Association of genomic subtypes of lower-grade gliomas with shape features automatically extracted by a deep learning algorithm". In: *Computers in Biology and Medicine* 109 (2019). DOI: 10.1016/j.compbiomed.2019.05.002.

[7] Mathilde Caron et al. "Deep clustering for unsupervised learning of visual features". In: *Proceedings of the European conference on computer vision (ECCV)*. 2018, pp. 132–149.

[8] Jia Deng et al. "ImageNet: A large-scale hierarchical image database". In: *2009 IEEE Conference on Computer Vision and Pattern Recognition*. 2009, pp. 248–255. DOI: 10.1109/CVPR.2009.5206848.

[9] Brittany N Dugger and Dennis W Dickson. "Pathology of neurodegenerative diseases". In: *Cold Spring Harbor perspectives in biology* 9.7 (2017), a028035.

[10] Martin Ester et al. "A density-based algorithm for discovering clusters in large spatial databases with noise". In: *kdd*. Vol. 96. 34. 1996, pp. 226–231.

[11] Mark Everingham and John Winn. "The PASCAL visual object classes challenge 2012 (VOC2012) development kit". In: *Pattern Anal. Stat. Model. Comput. Learn., Tech. Rep* 2007.1-45 (2012), p. 5.

[12] Navid Farahani, Anil V Parwani, Liron Pantanowitz, et al. "Whole slide imaging in pathology: advantages, limitations, and emerging perspectives". In: *Pathol Lab Med Int* 7.23-33 (2015), p. 4321.

[13] John C Fiala. "Mechanisms of amyloid plaque pathogenesis". In: *Acta neuropathologica* 114 (2007), pp. 551–571.

[14] Vivek Gopal Ramaswamy et al. "A Scalable High Throughput Fully Automated Pipeline for the Quantification of Amyloid Pathology in Alzheimer's Disease Using Deep Learning Algorithms". In: *bioRxiv* (2023), pp. 2023–05.

[15] Patricia Fierro Hernández. "Quantitative amyloid-β measures associate with cognitive performance in self-reported cognitively healthy centenarians". July 2022.

[16] *Geometrical product specifications (GPS) — Geometrical tolerancing — Tolerances of form, orientation, location and run-out.* Standard. Geneva, CH: International Organization for Standardization, Feb. 2017.

[17] Itseez. *Open Source Computer Vision Library.* https://github.com/itseez/opencv. 2015.

[18] Norihiko Iwamoto et al. "Distribution of amyloid deposits in the cerebral white matter of the Alzheimer's disease brain: relationship to blood vessels". In: *Acta neuropathologica* 93 (1997), pp. 334–340.

[19] Alex Krizhevsky. "One weird trick for parallelizing convolutional neural networks". In: *arXiv preprint arXiv:1404.5997* (2014).

[20] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. "Imagenet classification with deep convolutional neural networks". In: *Advances in neural information processing systems* 25 (2012).

[21] Zhengfeng Lai et al. "BrainSec: Automated Brain Tissue Segmentation Pipeline for Scalable Neuropathological Analysis". In: *IEEE Access* 10 (2022), pp. 49064–49079. DOI: 10.1109/access.2022.3171927. URL: https://doi.org/10.1109/access.2022.3171927.

[22] Tsung-Yi Lin et al. "Microsoft coco: Common objects in context". In: *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13.* Springer. 2014, pp. 740–755.

[23] Sharada P Mohanty, David P Hughes, and Marcel Salathé. "Using deep learning for image-based plant disease detection". In: *Frontiers in plant science* 7 (2016), p. 1419.

[24] Muhammad Khalid Khan Niazi, Anil V Parwani, and Metin N Gurcan. "Digital pathology and artificial intelligence". In: *The lancet oncology* 20.5 (2019), e253–e261.

[25] Kay R. J. Oskal et al. "A U-net based approach to epidermal tissue segmentation in whole slide histopathological images". In: *SN Applied Sciences* 1.7 (June 2019). DOI: 10.1007/s42452-019-0694-y. URL: https://doi.org/10.1007/s42452-019-0694-y.

[26] Nobuyuki Otsu. "A Threshold Selection Method from Gray-Level Histograms". In: *IEEE Transactions on Systems, Man, and Cybernetics* 9.1 (1979), pp. 62–66. DOI: 10.1109/TSMC.1979.4310076.

[27] Adam Paszke et al. "PyTorch: An Imperative Style, High-Performance Deep Learning Library". In: *Advances in Neural Information Processing Systems 32.* Curran Associates, Inc., 2019, pp. 8024–8035. URL: http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf.

[28] F. Pedregosa et al. "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

[29] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. "U-net: Convolutional networks for biomedical image segmentation". In: *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*. Springer. 2015, pp. 234–241.

[30] A. C. Ruifrok and D. A. Johnston. "Quantification of histochemical staining by color deconvolution". In: *Anal Quant Cytol Histol* 23.4 (Aug. 2001), pp. 291–299.

[31] Terri-Leigh Stephen et al. "Machine Learning Classification of Alzheimer's Disease Pathology Reveals Diffuse Amyloid as a Major Predictor of Cognitive Impairment in Human Hippocampal Subregions". In: *bioRxiv* (2023), pp. 2023–05.

[32] Ziqi Tang et al. "Interpretable classification of Alzheimer's disease pathologies with a convolutional neural network pipeline". In: *Nature Communications* 10.1 (May 2019). DOI: 10.1038/s41467-019-10212-1. URL: https://doi.org/10.1038/s41467-019-10212-1.

[33] Lary C Walker. "Aβ plaques". en. In: *Free Neuropathology* (2020), Vol 1 (2020). DOI: 10.17879/FREENEUROPATHOLOGY-2020-3025. URL: https://www.uni-muenster.de/Ejournals/index.php/fnp/article/view/3025.

[34] Stéfan van der Walt et al. "scikit-image: image processing in Python". In: *PeerJ* 2 (June 2014), e453. ISSN: 2167-8359. DOI: 10.7717/peerj.453. URL: https://doi.org/10.7717/peerj.453.

[35] Anderson M. Winkler et al. "Cortical thickness or grey matter volume? The importance of selecting the phenotype for imaging genetics studies". In: *NeuroImage* 53.3 (2010). Imaging Genetics, pp. 1135–1146. ISSN: 1053-8119. DOI: https://doi.org/10.1016/j.neuroimage.2009.12.028. URL: https://www.sciencedirect.com/science/article/pii/S1053811909013160.

[36] HM Wisniewski et al. "Spectrum of morphological appearance of amyloid deposits in Alzheimer's disease". In: *Acta neuropathologica* 78.4 (1989), pp. 337–347.