Quantification of Amyloid-Beta Plaques Using Few-Shot Learning

Master Thesis

Janaína Moreira-Kanaley

Delft University of Technology

TU Delft

Quantification of Amyloid-Beta Plaques Using Few-Shot Learning

by

Janaína Moreira-Kanaley

Student Name	Student Number
Janaína Moreira-Kanaley	5027594

Project Duration:	November, 2023 - August, 2024
Faculty:	Faculty of Electrical Engineering, Mathematics and Computer Science, Delft
Thesis Committee:	Prof. dr. ir. M. J. T. Reinders, TU Delft
	Prof. dr. ir. K. A. Hildebrandt, TU Delft
	Dr. ir. X. Zhang, TU Delft
	Dr. J. J. M. Hoozemans, Amsterdam UMC
	S. K. Rohde, Amsterdam UMC

Cover: Brain Background by pikisuperstar on Freepik
Style: TU Delft Report Style, with modifications by Daan Zwaneveld

TU Delft

Abstract

Alzheimer's disease (AD) is a neurodegenerative disorder prevalent in older adults, leading to loss in memory, cognitive, and executive function. A characteristic feature of AD is the accumulation of amyloid-beta (Aβ) plaques, which are extracellular deposits of Aβ protein primarily found in the grey matter. These Aβ deposits can appear in different forms. The six primary Aβ deposits covered in this work are: diffuse plaques, cored plaques, compact plaques, coarse grained plaques, cerebral amyloid angiopathy (CAA), and subpial deposits. It is speculated that some of the plaques types may be more harmful than others. Given that AD primarily affects an older population, the question arises of how individuals with AD differentiate from cognitively healthy centenarians (people over 100 years old). Consequently, this work attempts to classify and analyse different Aβ types present in donated brain tissue of cognitively healthy centenarians who escaped dementia and individuals diagnosed with AD. The main goal is to identify differences between these two cohorts. However, a challenge in this process is that the brain tissues contain many Aβ plaques, making manual identification difficult in terms of time and labor. To address this, a fine-tuned ResNet50 Aβ plaque classifier was developed in this research that was integrated into an Aβ detection pipeline capable of localising plaques. The model was initially pre-trained on the ImageNet dataset through contrastive learning, and subsequently fine-tuned using few-shot learning with a small number of annotated samples (315). The annotations included the six primary Aβ types whose structure are known and well-defined, and three other anomaly Aβ types that served to filter out irregular plaques in the unlabeled Aβ data. After performing 5-fold cross-validation, the fine-tuned models demonstrated an average accuracy of 85.71% and precision of 89.47% on the primary types. The final classifier used on the unlabeled Aβ dataset was an ensemble model that incorporated majority voting, combining the predictions of the five models trained during cross-validation. Aβ loads were calculated for each primary Aβ type based on the classifier's predictions. It was observed that across all considered primary Aβ types, the centenarians' Aβ loads were statistically significantly lower compared to the AD cohort. The lower Aβ load in centenarians also held true across the frontal, temporal, parietal, and occipital cerebral regions for each primary Aβ type. To partially validate the model's performance, correlations were computed between the predicted Aβ loads of the primary Aβ types and neuropathological assessments collected from 75 centenarians. These assessments are common in related works and serve as a benchmark for the ensemble model. They include the Thal Aβ phase, which categorizes the distribution of general Aβ in the brain; the Thal CAA stage, which measures the severity of CAA; and CERAD NP scores, which evaluates the spread of neuritic plaques in the brain, which are a subset of cored plaques. Consequently, statistically significant positive correlations were revealed between: the Thal Aβ phase and the Aβ load of all primary types (r ranging 0.59-0.73); the Thal CAA stage and Aβ load of predicted CAA deposits ($r = 0.66$); and the CERAD NP scores and Aβ load of cored plaques ($r = 0.68$). Since the correlated types coincide with what the benchmark staging schemes measure, the model's predictions seems to align with existing literature.

Contents

1 Introduction

Alzheimer's disease (AD) is a neurodegenerative disorder that predominantly occurs in older adults, associated with significant impairments in memory, cognition, and executive function. It is the leading cause of dementia worldwide [1]. A pathological hallmark of Alzheimer's is the accumulation of amyloid-beta (Aβ) plaques, which are extracellular deposits of the Aβ protein mainly located in the grey matter of the brain [2]. This study specifically investigates six different types of Aβ depositions: diffuse plaques, cored plaques, compact plaques, coarse-grained plaques, cerebral amyloid angiopathy (CAA), and subpial aggregation. It is speculated that Aβ plaques begin as diffuse plaques, then evolve to cored plaques, and finally become compact plaques [3]. However, despite ongoing research, the precise role these deposits play in the progression of AD remains unclear. A detailed analysis involving the categorization and quantification of different Aβ formations could help provide more insights to the physiology of AD and possible targets for treatment. Considering the large quantity of Aβ plaques in donated brain samples, manually annotating these deposits is both labor intensive and time consuming. This paper aims to automate the annotation of different plaques types by exploring the use of a fine-tuned ResNet50 model for the classification of plaques in a supervised few-shot learning context.

Considering that Alzheimer's disease is associated with an aging population, it may be useful to examine what differentiates very old individuals who do not develop AD. Consequently, this research compares Aβ plaque classifications between two different groups: cognitively healthy centenarians (people aged 100 years or older) participating in the 100-plus (100+) study at Amsterdam University Medical Center (AUMC) and individuals diagnosed with AD based on meeting neuropathological criteria during life. The 100+ study assesses cognitive health in centenarians through neuropsychological evaluations, while also collecting genetic and brain tissue data upon their passing [4]. Previous research indicates that Aβ plaques are present in the majority of centenarians and suggests that healthy old individuals may possess a resilience to these pathologies [5]. Additionally, other studies have proposed genetic factors, indicating that healthy centenarians possess alleles which protect them against the presence of Aβ plaques [6]. These findings can be expanded by examining, outside of genetic factors, whether the cognitively healthy centenarians show different distributions of certain Aβ plaque types compared to the AD population. A convolutional neural network (CNN) model that categorizes Aβ plaques would facilitate an analysis between these two groups.

Several studies have developed pipelines for detecting and classifying a subset of the investigated Aβ deposits. Tang et al. [7] introduced an end-to-end CNN-based pipeline that processes whole slide images and categorises Aβ pathologies into cored plaques, diffuse plaques, and CAA. Similarly, [8] utilised a Mask R-CNN for the analysis of post-mortem tissue, achieving human-comparable accuracy in classifying the same plaque types. Stephen et al. [9] trained a random forest classifier to distinguish between dense and diffuse plaques, highlighting diffuse amyloid pathology as a major driver of AD in their cohort. While these works have successfully detected and classified Aβ plaques, they focus on a subset of the plaque types covered. This work explores a broader range of six distinct Aβ deposit types, allowing for a more in-depth analysis of associations.

In contrast to literature that relies on tens of thousands of plaque annotations [7], the model used in this research employs only 315 annotations (35 per Aβ type), which reduces the costs associated with manual annotations and potentially simplifies the process of incorporating new plaque types into the classification model.

In addition to the previously discussed benefits, this work expands on an existing pipeline that handles the classification of Aβ plaques. Chiel de Vries' thesis pipeline involves segmenting grey matter from whole slide images, localising plaques within the grey matter, and using unsupervised clustering to categorise the plaques [10]. However, the categorisation stage of the pipeline yielded unsatisfactory performance in correctly classifying Aβ deposits. Consequently, this research employs the pipeline from [10] and improves its Aβ categorization by incorporating a supervised approach.

Building on the foundation of previous research, this study investigates how to reliably categorise Aβ plaques with a small amount of manually annotated data. The process involves locating Aβ plaques in segmented grey matter [10], as described in sections 2.1 and 2.2. Annotated Aβ plaques are then used to fine-tune a ResNet50 model, which was originally pre-trained on the ImageNet dataset through contrastive learning. The methods for setting up and training the ResNet50 model are discussed in Section 2.4. Following this, the model's performance is assessed through quantitative metrics in Section 3.1.2 and visualisations of Aβ classifications in Section 3.1.3. Subsequently, Section 3.2.1 compares the distributions of the Aβ type predictions made for the 100+ and AD cohorts to identify any significant differences between them. Additionally, Section 3.2.2 shows the correlations between the Aβ loads for each predicted primary type, and cognitive and neuropathological assessment data collected from centenarians. A reflection of the results is discussed in Chapter 4. The main conclusions from the research can be found in Chapter 5.

2 Methodology

2.1. Datasets

2.1.1. Whole Slide Images for 100+ and AD Cohorts

The Aβ plaques analysed in this research are extracted from brain slices collected from both healthy centenarians and individuals diagnosed with AD. The brain slices gathered in the 100+ study, led by AUMC, originate from Dutch centenarians who self-reported as cognitively healthy [4]. The study gathered extensive data, including demographics, life history, medical history, genealogy, neuropsychological assessments, blood samples, and post-mortem brain donations. These donated brain samples were scanned using the Olympus VS200 [11] and saved as high-resolution whole slide images (WSIs) in red-green-blue (RGB). Figure 2.1 shows an example of a WSI. For this work specifically, brain samples were considered from 93 individuals in the 100+ cohort and 29 individuals in the AD cohort. Each WSI represents a slice taken from one of four cerebral regions: the middle frontal gyrus, inferior parietal lobule, temporal pole, and occipital pole. Slides are stained immunohistochemically for the identification of Aβ deposits, as described in [11]. Further details about the WSI dataset can be found in Table 2.1. Some of the centenarians have had Aβ loads pre-computed for the four brain slice regions using a pixel level classifier [11]. The Aβ load represents the percentage of grey matter area covered with Aβ positivity.

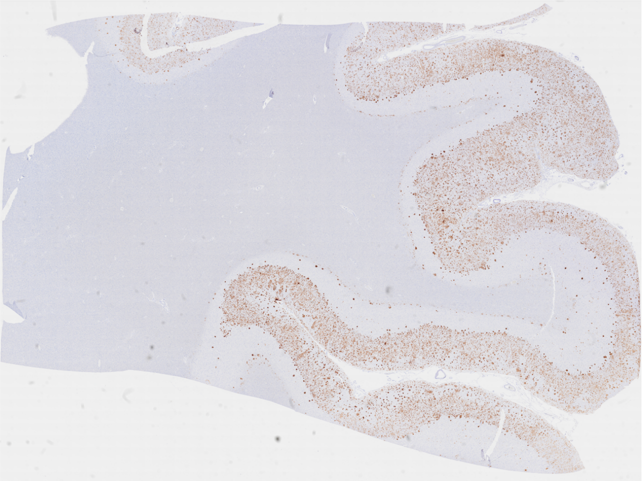


Figure 2.1: A brain slice stained immunohistochemically and saved as a WSI. Aβ plaques are visible as small brown stains in the grey matter.

Table 2.1: WSI statistics on the 100+ and AD data. Image dimensions are given as mean accompanied by the standard deviation.

	100+	AD
Number of slides	365	116
Number of individuals	93	29
Image height (pixels)	73,206 ± 8,670	50,618 ± 30,513
Image width (pixels)	96,611 ± 9,079	68,940 ± 46,701
Image height (μm)	20,058 ± 2,376	13,869 ± 8,361
Image width (μm)	26,471 ± 2,488	18,890 ± 12,796
Frontal slides	92	29
Temporal slides	89	29
Parietal slides	92	29
Occipital slides	92	29

2.1.2. Cognitive and Neuropathological Assessments for 100+ Cohort

The 100+ study includes data from various sources in addition to the WSIs. Among these sources, centenarians completed five tests (see Table 2.2) aimed at evaluating their cognitive function: the Mini-Mental State Examination (MMSE), Digit Span Backward (DSB), Digit Span Forward (DSF), Key Search (KS), and the Clock Drawing Test (CDT). Furthermore, neuropathological staging data was collected for some centenarians, which includes the Thal Aβ phase, Thal CAA stage, and Consortium to Establish a Registry for Alzheimer's Disease Neuritic Plaque (CERAD NP) scores. The Thal Aβ phase assesses the spread of Aβ deposition to different areas of the brain [12], which is categorised into six phases depicted in Table 2.3. The Thal CAA stage evaluates the distribution of CAA [13], with its stages described in Table 2.4. The CERAD NP score measures the density of neuritic plaques, a subset of cored plaques, in the brain [14]. Its stages are outlined in Table 2.5. Together, these different measures can capture an evaluation of neurodegeneration from multiple perspectives. In addition, the neuropathological stages are considered the gold standard for neuropathological evaluation in the field, which can serve as a benchmark for assessing the model.

Table 2.2: Descriptions of cognitive tests used in the 100+ Study. Descriptions are based on Holstege et al. [4].

Test	Description
Mini-Mental State Examination	A 30-point questionnaire that measures cognitive impairment.
Digit Span Backward	Tests working memory by asking participants to repeat a series of digits in reverse order.
Digit Span Forward	The participant repeats a series of digits in the same order.
Key Search	The participant draws a path through a field to search for an imagined set of lost keys.
Clock Drawing Test	Participants are asked to draw numbers on the face of a clock and fill in a specific time.

Table 2.3: A summary of the Thal Aβ phases [12].

Thal Aβ Phase	Description
Phase 0	No Aβ deposition.
Phase 1	Aβ deposits are found exclusively in the neocortex.
Phase 2	Additional involvement of the allocortex.
Phase 3	Aβ deposits appear in the diencephalic nuclei, striatum, and cholinergic nuclei of the basal forebrain.
Phase 4	Several brainstem nuclei exhibit Aβ deposits.
Phase 5	Aβ deposition extends to the cerebellum.

Table 2.4: Thal CAA stages and their descriptions [13]. CAAs are Aβ deposits within the blood vessel walls of the brain.

Thal CAA Stage	Description
Stage 0	No CAA found.
Stage 1	CAA found in leptomeningeal and neocortical vessels.
Stage 2	CAA extends to allocortical and midbrain vessels.
Stage 3	Aβ deposition is observed in larger cortical vessels, such as the basal ganglia, thalamus, and lower brainstem.

Table 2.5: CERAD NP stages and their descriptions [14]. Neuritic plaques are a subset of cored plaques.

CERAD NP Stage	Description
Stage 0	None: No neuritic plaques observed.
Stage 1	Sparse: Few neuritic plaques observed.
Stage 2	Moderate: A moderate density of neuritic plaques present.
Stage 3	Frequent: A high number of neuritic plaques present.

2.2. Pipeline

A machine learning pipeline is used for processing the brain slice WSIs to extract and classify Aβ plaques. It is based on previous work on this topic [10]. The pipeline consists of three main steps: segmentation of the grey matter, localisation of Aβ plaques within the grey matter, and classification of the located Aβ plaques. An overview of the steps are depicted as different colored stages in Figure 2.2, where the final categorisation stage is the focus of this work. The first two stages of the pipeline are previously completed work of [10].

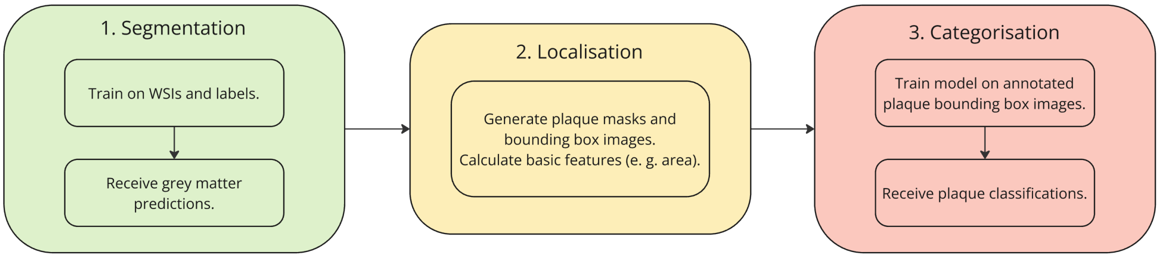
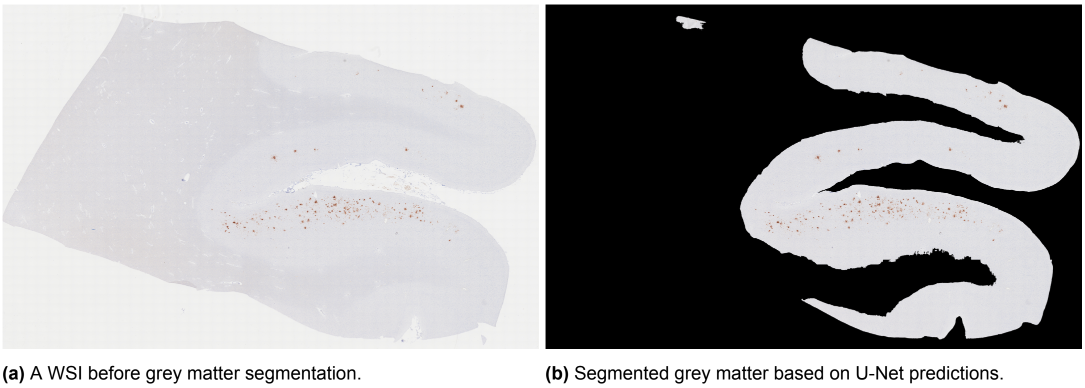

1. Segmentation
Train on WSIs and labels.
Receive grey matter predictions.
2. Localisation
Generate plaque masks and bounding box images.
Calculate basic features (e. g. area).
3. Categorisation
Train model on annotated plaque bounding box images.
Receive plaque classifications.

Figure 2.2: The overall structure of the pipeline responsible for locating and classifying the plaques in WSIs. The three stages of the pipeline are depicted in boxes of different colors. The smaller sub-boxes indicate steps within a stage.

2.2.1. Grey Matter Segmentation

In the segmentation stage, the pipeline separates grey matter from the rest of the brain tissue. This process removes areas within the WSIs that do not contain Aβ plaques, thereby saving on computational time. Figure 2.3a displays an example of a WSI, while Figure 2.3b shows the results after it undergoes segmentation. The segmentation is performed using a CNN model known as U-Net, which is trained on pixel-level annotations of grey matter provided by experts. To manage the high resolution of the images and reduce required computational resources, training is conducted in smaller patches.



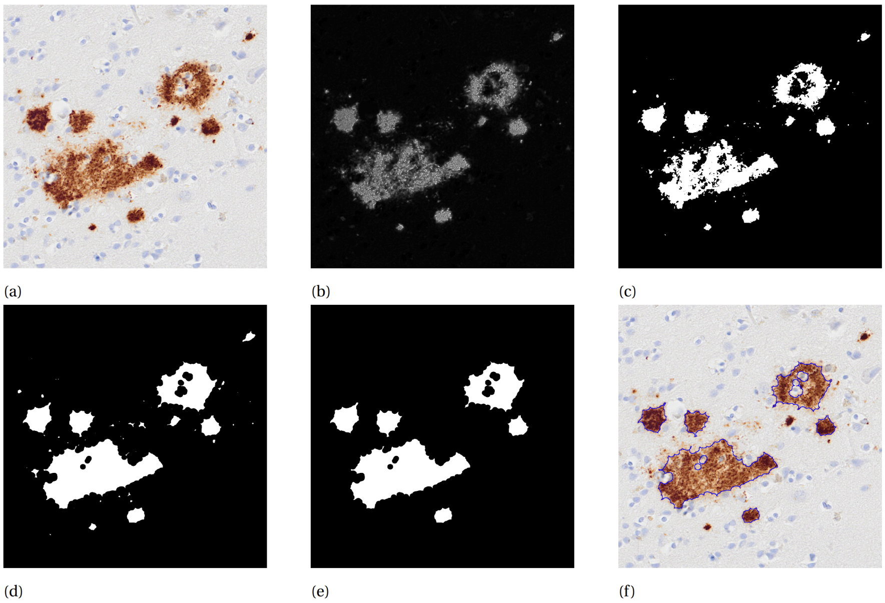
(a) A WSI before grey matter segmentation.

(b) Segmented grey matter based on U-Net predictions.

Figure 2.3: A visualisation of the grey matter segmentation process. (a) An example WSI given as input. (b) The result after applying the grey matter segmentation mask predicted by the U-Net model on the WSI.

2.2.2. Plaque Localisation

The localisation stage locates plaques in the segmented grey matter from the segmentation stage. Initially, it transforms WSI patches from RGB into the hematoxylin-eosin-DAB (HED) color space. The Aβ plaques are specifically stained with 3,3'-Diaminobenzidine (DAB), therefore only the DAB channel is extracted from HED. Subsequently, Otsu's threshold selection algorithm is applied to create a binary mask of the image, the result of which can be seen in Figure 2.4. Finally, connected components in the mask are identified and filtered based on size. Any plaques whose masked area is lower than the area of a circle with a diameter of 10µm is deemed pathologically insignificant in this step. Components of significant size that are disconnected to each other are viewed as separate plaques.



(a)

(b)

(c)

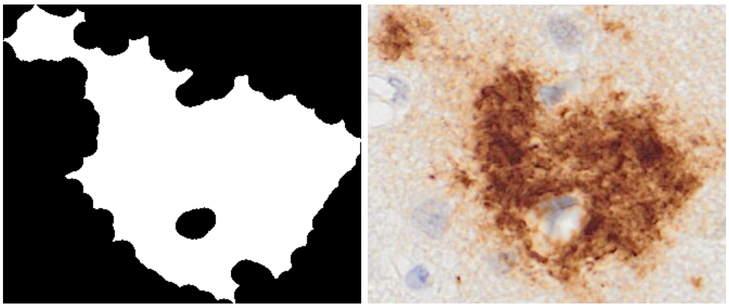
(d)

(e)

(f)

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" [10].

After locating the plaques, relevant information is recorded, including the RGB bounding box image of each plaque and its corresponding binary mask. Additionally, basic features such as plaque area are computed. An example of an extracted plaque from the localisation algorithm can be seen in Figure 2.5.



(a) Binary mask of a located plaque. **(b)** Bounding box of a located plaque.

Figure 2.5: Example of a located plaque from the localisation algorithm. (a) The binary mask of a plaque. (b) The corresponding bounding box of the plaque.

2.2.3. Plaque Categorisation

The categorisation stage handles the classification of Aβ plaques into different types. These plaques can be accessed as bounding box RGB images, which were generated during the localisation stage of the pipeline. In the categorisation stage, the model is trained using annotated Aβ images and applied to generate predictions. The focus of the research is on this stage, with details given in further sections.

2.3. Annotated Aβ Types

This section describes the Aβ types considered in the research. For the purpose of supervised learning, 315 Aβ plaques were annotated by pathology expert S. K. Rohde, with 35 annotations made for each Aβ type. Descriptive statistics on the number of plaques extracted during the localisation stage and their image sizes are provided in Table 2.6.

Table 2.6: Descriptive statistics on the Aβ plaque images extracted from the WSIs through the Aβ detection pipeline of [10]. Image sizes are given as mean ± standard deviation.

	100+	**AD**
Number of plaques	1,873,831	2,271,910
Plaque height (pixels)	100 ± 83	109 ± 91
Plaque width (pixels)	104 ± 90	114 ± 95
Plaque height (µm)	27 ± 23	30 ± 25
Plaque width (µm)	28 ± 25	31 ± 26

2.3.1. Primary Aβ Types

The model is trained to recognize in total nine types of Aβ plaques, but the primary focus is on six main types: diffuse plaques, cored plaques, compact plaques, coarse-grained plaques, CAA, and subpial depositions. These six primary plaques have well-defined and well-documented forms. In this research, the primary plaques are the ones involved in analyses of Aβ plaque distributions between the 100+ and AD cohorts. Figure 2.6 depicts annotated examples of each Aβ deposit type. More detailed descriptions of the six primary Aβ plaques are provided below:

- *Diffuse plaques* are loosely arranged structures of Aβ deposits with irregular and ill-defined margins [15].
- *Cored plaques* have a dense compact core often surrounded by a less dense, diffuse corona [16].
- *Compact plaques* consists of a core with a dense accumulation of Aβ deposits, without a corona [16].
- *Coarse grained plaques* have a structure of multiple cores and Aβ-devoid pores [17].
- *Cerebral Amyloid Angiopathy (CAA)* is characterized by Aβ deposits in the walls of the cerebral blood vessels and has a circular appearance. CAA in this work specifically addresses CAA type 2, which features Aβ deposits in leptomeningeal and cortical vessels, excluding cortical capillaries [18].

- *Subpial depositions* is the accumulation of Aβ beneath the pia matter and appears in a band-like shape [16].

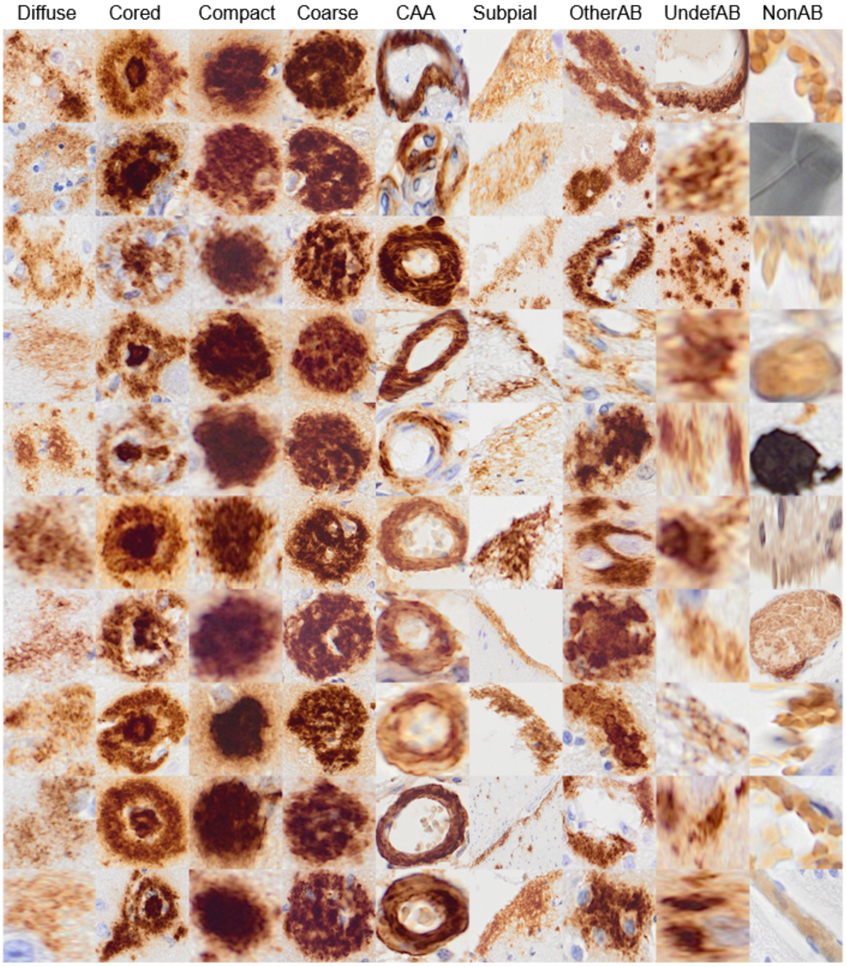


Figure 2.6: Samples of annotated plaques where each column shows which Aβ type they belong to.

2.3.2. Anomaly Aβ Types

While the Aβ plaque dataset is assumed to contain plaques that can be categorised into the six primary types, it may also contain other deposit formations that do not fit into these categories. To minimize false positive classifications, the model is trained to recognize and filter out these unconventional anomalous plaques. The anomalous plaques may also include plaques that are not covered by the model but do belong to other categories, such as CAA type 1 [18], cotton wool plaques [19], or dysphoric amyloid angiopathy [20]. Ultimately, the problem addresses a total of nine classes, including the six primary plaque types and three categories of anomaly plaques. The three anomaly plaque types can be seen in Figure 2.6 and are described below:

- *Other amyloid-beta plaques (OtherAB)* are plaques that have a well-defined bounding box image but do not fit into any of the six primary plaque descriptions. They may fit into another Aβ type that is not considered by the model.
- *Undefined amyloid-beta plaques (UndefAB)* are plaques that show Aβ staining, but their bounding box image is too ambiguous (e.g., due to blur or cropping) to be confidently categorized.
- *Non amyloid-beta (NonAB)* are bounding box images that do not contain any Aβ markers. They are false positive detections from the localisation stage of the pipeline. The images can include, for example, a piece of glass mistakenly included in the plaque localisation process.

2.4. Few-Shot Learning

2.4.1. Model

A 50-layered residual network (ResNet50) architecture from [21] is employed for the classification of Aβ plaques. ResNet has many applications in image recognition tasks [22], [23] and is used in this research for transfer learning. ResNet50 is initially pre-trained on the ImageNet dataset [24], which consists of images from 1000 different classes, and subsequently fine-tuned on the annotated plaques.

Several contrastive learning models use ResNet50 as a base encoder to effectively learn unsupervised representations for different tasks [25], [26], [27]. Some of these models also manage to successfully fine-tune ResNet on a small percentage of annotated data [25]. For instance, the Simple Contrastive Learning (SimCLR) model achieves 75.5% top-5 accuracy on ImageNet when fine-tuning ResNet50 on only 1% of labels [25]. Building on the success of other contrastive learning models, the ResNet50 encoder from the Momentum Contrast (MoCo) model pre-trained on ImageNet is selected for fine-tuning on the annotated plaques [26].

Architecture

ResNet50 consists of 50 layers structured with residual blocks [21], whose architecture can be seen in Figure 2.7. It addresses the vanishing gradient problem by allowing gradients to pass through skip connections. The residual blocks consist of convolutional layers, batch normalization, and ReLU activations. The original model ends with an average pooling layer and a 1000-dimensional output fully-connected (FC) layer. To customize the model for classifying plaques, the pre-existing FC layer is replaced with a new FC with an output dimension of nine. The dimension corresponds to the nine classes that have been established. A dropout layer is also attached before the last layer as a regularisation technique to reduce overfitting. Softmax is used as the activation function for the final layer to generate class probabilities.

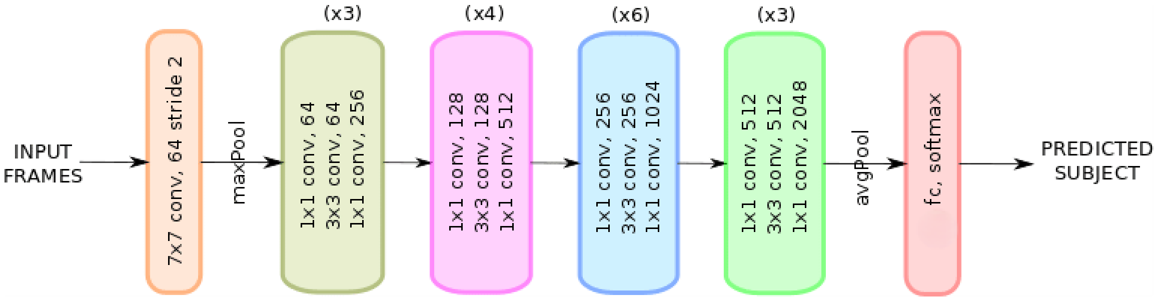


Figure 2.7: Visualisation of the architecture of the ResNet50 model. Original image is taken from [28]. The last FC layer of the model has been replaced by another with an output dimension of nine. This adjustment allows the model to classify the nine Aβ plaque types.

Input images are resized to 128x128 pixels. Although downsampling might result in some loss of detail, this resolution is still above the average plaque image sizes. This resizing can help mitigate overfitting, which is important considering the limited annotations.

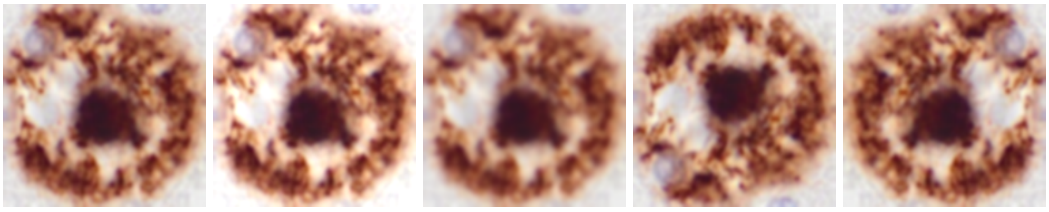
Classification Rule

To classify plaques based on the nine output class probabilities, a thresholding method is used. The Receiver Operating Characteristic (ROC) curve [29] is calculated for each class versus the rest, which plots for each class the true positive rate against the false positive rate of the model at different classification thresholds. Before selecting the thresholds, predictions are first made by assigning the class with the highest predicted probability score. Thresholds for each class are then determined based on trying to optimize the balance between a higher true positive rate and a lower false positive rate. Once this threshold is set, the classification method is changed so that plaques are assigned to the class with the higher probability only if its probability is equal to or exceeds the threshold. If no class meets this condition, the class with the highest probability is chosen. Classification thresholds used can be found in Appendix A.

The final classification model comprises of an ensemble model. This model approach takes advantage of more data samples, while reducing the instability of using only a single model trained on a small subset of data. The ensemble consists of five models trained with 5-fold cross-validation. For each of the five models, class probability thresholds are determined using predictions from the corresponding validation set in cross-validation. The final ensemble model operates using hard voting: each of the five models vote for the class label of an input plaque, and the class with the majority of votes is assigned to the plaque.

2.4.2. Data Augmentation

Data augmentation plays a crucial role in expanding the effective dataset size and introducing variability to the limited number of annotated samples. Consequently, several augmentation techniques are applied on the training data before training the model. In addition to the original image, each plaque sample also undergoes transformations through four operations: horizontal flip, vertical flip, color jitter, and Gaussian blur. The impact of these transformations and specific parameters used are shown in Figure 2.8 and Table 2.7, respectively. After applying these augmentations, the dataset is effectively multiplied by a factor of five. The augmentations were chosen based on slight modifications, avoiding other operations such as cropping to ensure the plaques remain identifiable and retain information that may be critical for correct classification.



(a) Original input image **(b)** Color Jitter **(c)** Guassian Blur **(d)** Horizontal Flip **(e)** Vertical Flip

Figure 2.8: Overview of data augmentation techniques used. (a) The original input image given to the model. (b) Brightness, saturation, contrast, and hue are randomly adjusted for the image. (c) Detail and noise in the image are reduced by smoothing with a Guassian filter. (d) The image is flipped on the vertical axis. (e) The image is flipped on the horizontal axis.

Table 2.7: Overview of the parameters for the employed data augmentation techniques.

Augmentation Technique	Parameters
Color Jitter	Brightness: 0.2, Contrast: 0.2, Saturation: 0.2, Hue: 0.0
Gaussian Blur	Kernel Size: 7, Sigma Range: [0.1, 2.0]
Horizontal Flip	-
Vertical Flip	-

2.4.3. Training

To obtain an ensemble model and achieve a more stable classification method, 5-fold cross-validation is employed, which splits the annotated dataset into five groups. Considering that the dataset consists of 35 labeled samples per class, each group contains seven labels per class, totaling 63 samples per group. The ensemble model is built from five models trained on different configurations of these folds. For each model, one group serves as the test set, while the remaining four groups are used for training and validation. From these four groups, the training set takes 23 labels per class, and the validation set takes five labels per class. Table 2.8 provides an overview of the parameters used for each fold configuration.

The hyperparameters for batch size, dropout probability, learning rate, and weight decay were selected by hyperparameter tuning using PyTorch Optuna [30]. Optuna uses the Tree-structured Parzen Estimator (TPE) approach, which models the objective function and chooses hyperparameters that have the highest potential to improve the objective. In this case, the objective is based on minimizing the cross-entropy loss averaged over the 5-fold validation sets. Details on explored and chosen hyperparameters

Table 2.8: Number of annotated plaque samples used for each train, validation, and test splits of the dataset. This holds for every fold configuration during 5-fold cross-validation.

Data Split	Samples per Class	Total Samples across Classes	Total Samples After Augmentation
Train Set	23	207	1035
Validation Set	5	45	-
Test Set	7	63	-

can be seen in Table 2.9.

During training, the Adam optimizer is used in conjunction with cosine annealing to schedule the learning rate. Additionally, weight decay is applied with Adam to help prevent overfitting. The number of epochs is determined by an early stopping mechanism, which halts training if the validation loss does not improve after five epochs.

Table 2.9: Hyperparameters explored during hyperparameter tuning. Chosen hyperparameters were selected based on the lowest validation set loss averaged over the trained 5-fold models.

Hyperparameter	Explored Values	Chosen Value
Batch Size	32, 64	64
Dropout Probability	0.1, 0.2, 0.3	0.2
Learning Rate	1e-5, 1e-4	1e-5
Weight Decay	1e-5, 1e-4, 1e-3	1e-5

2.4.4. Model Performance Evaluation

Quantitative Evaluation

The performance is evaluated using a variety of quantitative metrics: accuracy, recall, precision, F1 score, and ROC area under the curve (AUC). These metrics are calculated by averaging the performance of the five models on their corresponding cross-validation test sets. The average is taken to achieve a more reliable estimate of the model and to reduce the instability in classifications that might arise from taking only one data configuration.

To provide additional insights into the classification results and show the distribution of class predictions, a confusion matrix is plotted alongside the metrics based on the aggregated predictions from the five test sets. The confusion matrix displays the distribution of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) classifications for each class. Table 2.10 provides an overview of these terms for a binary classification problem, where positive means belonging to the class of interest and negative means not belonging to that class.

Predicted Ground Truth	Positive	Negative
Positive	True Positive (TP)	False Negative (FN)
Negative	False Positive (FP)	True Negative (TN)

Table 2.10: Overview of the elements of a confusion matrix in a binary classification setting. Rows are the original ground truth labels, while columns represent the predicted labels assigned by the model.

Common metrics are used to evaluate the model. Accuracy measures the ratio of correctly predicted instances out of the total (see Equation 2.1) and provides a general indication of the model's performance. It works better on balanced datasets, which is the case in this research. Precision (see Equation 2.2) evaluates the reliability of positive predictions, with high values indicating lower false positives. On the other hand, recall (see Equation 2.3) measures how well the positive predictions capture all ground truth positive cases, with high values meaning lower false negatives. Recall is generally more informative on imbalanced datasets compared to accuracy, where missing positive instances can be very costly (e. g. missing a cancer diagnosis). The F1 score is the harmonic mean of precision and recall, as established in Equation 2.4, which provides a balance between these two measures. The ROC

AUC score is the area under the ROC curve and evaluates the model's ability to distinguish between classes across all classification thresholds. A higher AUC indicates better discriminative performance, while a value of 0.5 means that it is equivalent to random guessing.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{2.1}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{2.2}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{2.3}$$

$$\text{F1} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{2.4}$$

Quantitative Evaluation of Primary Aβ Types versus All Aβ Types

The quantitative evaluation of the models is approached from two perspectives. The first includes evaluating whether the models can perform well when only the plaques of interest are expected in the dataset. The performance here is calculated based on an adjusted test set that contains solely the annotated primary plaques. Additionally, predictions are changed to only consider probabilities from the six primary classes. In this way, the effect of enabling the anomaly classifications in the model can be compared to this evaluation.

The second perspective involves evaluating the models across all nine Aβ plaque types considered in the research, which consist of both primary and anomaly types. In this case, the aim is to test the model's performance on the more realistic and practical application, namely on the unlabeled dataset which may contain various plaque types besides the primary ones.

Plaque Visualisations

To provide a visual understanding of which plaques are being classified to what types, classifications are pictured for randomly sampled plaques from both the labeled and unlabeled Aβ datasets. These visualisations are presented as an image matrix, where each column consists of plaque images assigned by the model to the corresponding Aβ type. In the case of the labeled Aβ dataset, plaques that are incorrectly classified are annotated with blue text indicating their true class, while correctly classified plaques have no text annotations. For the unlabeled dataset, since the true labels are unknown, no blue text annotations can be expected for its visualisation.

2.4.5. Model Application

To identify patterns in the distributions of Aβ for each Aβ class, the ensemble model is applied to the entire Aβ dataset from both the 100+ and AD cohorts. Furthermore, the model's predicted distributions are compared to existing literature, providing partial validation of its performance by assessing whether results align with what is expected. Overall, the analyses are conducted on predicted: Aβ type frequency distributions, Aβ type load distributions over all and per cerebral region, and Aβ type load correlations with the 100+ assessment data. Besides the frequency distributions, the conducted analyses only involve the primary Aβ types as these are the ones that are known and well-defined in literature [16].

Statistical Tests

In evaluating differences between groups (e. g. 100+ and AD cohorts), the Mann-Whitney U test [31] is employed. The Mann-Whitney U test is a non-parametric test that assesses whether the differences between distributions are significant. Its null hypothesis states that the distributions of two populations are identical. Similarly, the alternative hypothesis is that the distributions are not identical. Ultimately, the null hypothesis is rejected for a p-value smaller than 0.05, meaning that differences in two distributions are statistically significant.

While analysing correlations, Spearman rank correlation [32] is used. This is also a non-parametric test which asses the strength and direction of monotonic relationships between variables. Its null hypothesis is that there is no monotonic relationship between two populations. Consequently, if p-value is smaller than 0.05 the null hypothesis is rejected, meaning that there is a monotonic relationship.

Descriptive terms used for Spearman rank correlations are based on a rule of thumb [33]:

- None: $0.0 \leq |r| < 0.1$
- Poor: $0.1 \leq |r| < 0.3$
- Fair: $0.3 \leq |r| < 0.6$
- Moderate: $0.6 \leq |r| < 0.8$
- Very Strong: $0.8 \leq |r| < 1.0$
- Perfect: $|r| = 1$

Aβ Frequency Distributions

To provide a clear overview of how the predicted classes are distributed across the entire Aβ dataset, bar plots displaying plaque frequencies are used. The plots include both absolute and relative frequencies for each Aβ class. One bar plot figure is presented for all Aβ types including anomalies, while another only for the primary types of interest.

Aβ Load Distributions

To conduct a more reliable analysis of Aβ distributions between the 100+ and AD cohorts across the different classes, Aβ loads are calculated. A reason for using Aβ loads instead of Aβ frequencies is that the number of Aβ plaques per WSI can be affected by the size of the brain tissue sample taken. Therefore, to account for variations in grey matter area, Aβ load is calculated as the ratio of the area covered by Aβ types to the total grey matter area (see Equation 2.5). Ultimately, the Aβ load metric is well-established and aligns with measurements used in literature, which have demonstrated strong associations between Aβ load, AD diagnosis and cognitive decline in centenarians [11].

$$\text{Class A}\beta\text{ Load (\%)} = \frac{\text{Area covered in A}\beta\text{ deposits of a predicted class}}{\text{Total grey matter area}} * 100 \tag{2.5}$$

The Aβ load distributions for 100+ and AD cohorts are given in the form of box or violin plots per primary class. In addition, cohort distributions are also analysed per class across different cerebral regions. The Mann-Whitney U test is used to calculate statistical significance for differences between 100+ and AD cohorts for a class or region. Statistically significant p-values are indicated by asteriks: $* = p < 0.05$, $** = p < 0.01$, $*** = p < 0.001$.

Correlations for 100+ Assessments

To analyse possible relationships and partially validate the performance of the model, the correlations between the predicted Aβ type loads and 100+ (cognitive and neuropathological) assessment data are analysed. The assessment data used belongs to 75 centenarians. Any Aβ loads computed for correlation analyses, correspond to the WSIs of these 75 centenarians. The correlations are plotted as correlation matrices for: the neuropathological staging schemes (Thal Aβ phase, Thal CAA stage, CERAD NP), the Aβ loads pre-computed with a pixel classifier for the four cerebral regions, and the cognitive tests completed by the centenarians (MMSE, DSB, DSF, KS, CDT). Statiscal significance is indicated in the correlation matrix.

3

Results

3.1. Model Performance

3.1.1. Quantitative Evaluation: Primary Aβ Types

The evaluation of the five cross-validation models on tests sets containing only the six primary Aβ types, reveals an average overall accuracy of 90.48% (see Table 3.1). All other performance metrics also exceed 90%. For the individual Aβ classes, diffuse has it lowest values for recall at 85.71%, with highest for precision at 100%. Cored plaques, despite having perfect precision at 100%, exhibit the lowest recall among all classes at 71.43%. Compact has all metrics above 92%, with the lowest being precision at 92.14%. The standard deviation for compact plaques is highest for precision (7.21%) and recall (7.82%), indicating instability across the models for this class. Coarse has the lowest precision among the classes at 72.89%, which drops the F1 score to 80.97%. The variability for coarse is greatest for recall with a standard deviation of 7.82%. CAA achieves perfect scores across all metrics (100%). While subpial also has perfect scores for recall, its performance is lower for precision at 87.50%.

Notable patterns can be observed when considering the class prediction distribution in the confusion matrix in Figure 3.1. While most diffuse plaques are correctly predicted (85.71%), 14.29% are incorrectly predicted as subpial. This suggests that the model may have some difficulty distinguishing between subpial structures and diffuse plaques with loosely defined margins. The cored class has the highest misclassification rate of all classes, with 28.57% predicted as coarse plaques. This issue is not completely unexpected as coarse plaques are known to be composed of multiple Aβ cores. Considering other types, the models misclassify 5.71% of compact plaques as coarse and 8.57% of coarse plaques as compact. On the other hand, both CAA and subpial classes have perfect recall, resulting in no misclassifications for plaques belonging to those classes.

Table 3.1: A summary of the average performance of the trained models (given in percentages) for test sets consisting of only Aβ primary types. The performance metrics consist of accuracy, precision, recall, F1 score, and ROC AUC. Each metric is reported with its mean ± standard deviation computed across five cross-validation configurations, on the corresponding test fold for each model. The overall metrics are computed considering all relevant classes.

Class	Accuracy	Precision	Recall	F1 Score	ROC AUC
Overall	90.48 ± 1.68	92.09 ± 1.58	90.48 ± 1.68	90.48 ± 1.58	98.20 ± 0.67
Diffuse	-	100.00 ± 0.00	85.71 ± 0.00	92.31 ± 0.00	99.27 ± 1.02
Cored	-	100.00 ± 0.00	71.43 ± 0.00	83.33 ± 0.00	93.39 ± 2.67
Compact	-	92.14 ± 7.21	94.29 ± 7.82	92.94 ± 5.07	99.67 ± 0.45
Coarse	-	72.89 ± 4.47	91.43 ± 7.82	80.97 ± 4.53	97.31 ± 1.40
CAA	-	100.00 ± 0.00	100.00 ± 0.00	100.00 ± 0.00	100.00 ± 0.00
Subpial	-	87.50 ± 0.00	100.00 ± 0.00	93.33 ± 0.00	99.59 ± 0.29

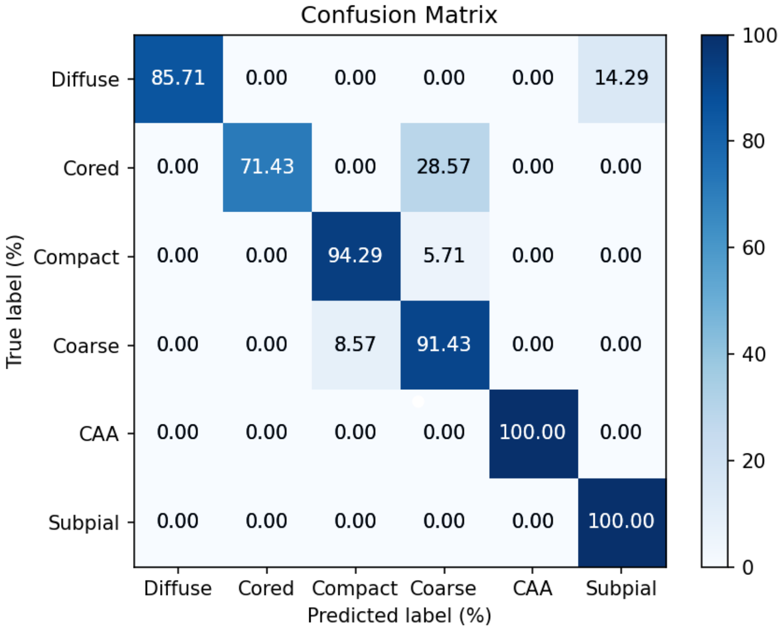


Figure 3.1: A confusion matrix depicting the overall classification performance of the five cross-validation models for test sets consisting of only Aβ primary types. Rows in the matrix represent the true class labels, while the columns represent the predicted classifications. In other words, each cell in the matrix presents the percentage of instances predicted by the model versus the true class labels. Values are computed by aggregating the predictions across the 5-fold cross-validation test sets for each of respective the models.

3.1.2. Quantitative Evaluation: Primary and Anomaly Aβ Types

When both primary and anomaly Aβ types are considered, there is a general decrease in the overall average performance for the primary types, as shown in Table 3.2 (includes anomaly types) compared to Table 3.1 (excludes anomaly types). Overall accuracy for primary drops from 90.48% to 85.71%, precision from 92.09% to 89.47%, recall from 90.48% to 85.71%, F1 score from 90.48% to 86.61%, and ROC AUC from 98.20% to 97.30%. The highest standard deviations for the primary types can be seen in diffuse with 11.41% for precision, compact with 12.78% for recall, and subpial with 11.95% for recall. Individual class comparisons between prediction distributions can be made by viewing the confusion matrix for all Aβ types in Figure 3.2 alongside Table 3.2.

Table 3.2: A summary of the average performance of the trained models across all considered Aβ classes is given in percentages. The performance metrics consist of accuracy, precision, recall, F1 score, and Receiver Operating Characteristic Area Under the Curve (ROC AUC). Each metric is reported with its mean value ± standard deviation computed across five cross-validation configurations, on the corresponding test fold for each model. The overall metrics are computed based on all nine classes.

Class	Accuracy	Precision	Recall	F1 Score	ROC AUC
Overall	76.83 ± 1.81	82.49 ± 2.39	76.83 ± 1.81	77.00 ± 1.51	95.84 ± 0.40
Overall (Primary)	85.71 ± 2.38	89.47 ± 3.24	85.71 ± 2.38	86.61 ± 0.71	97.30 ± 0.75
Overall (Anomaly)	59.05 ± 2.61	68.54 ± 9.80	59.05 ± 2.61	57.79 ± 5.23	92.91 ± 1.32
Diffuse	-	92.14 ± 11.41	82.86 ± 6.39	86.73 ± 5.48	97.81 ± 1.52
Cored	-	100.00 ± 0.00	68.57 ± 6.39	81.21 ± 4.74	89.64 ± 2.72
Compact	-	92.14 ± 7.21	91.43 ± 12.78	91.14 ± 6.68	99.80 ± 0.28
Coarse	-	74.00 ± 2.24	88.57 ± 6.39	80.47 ± 1.05	97.45 ± 2.06
CAA	-	97.50 ± 5.59	100.00 ± 0.00	98.67 ± 2.98	100.00 ± 0.00
Subpial	-	81.03 ± 4.46	82.86 ± 11.95	81.41 ± 4.95	99.13 ± 0.29
OtherAB	-	74.22 ± 20.46	57.14 ± 0.00	63.52 ± 8.53	93.32 ± 2.54
UndefAB	-	83.33 ± 23.57	40.00 ± 11.95	51.56 ± 9.08	94.23 ± 1.03
NonAB	-	48.06 ± 11.51	80.00 ± 12.78	58.28 ± 4.71	91.17 ± 4.02

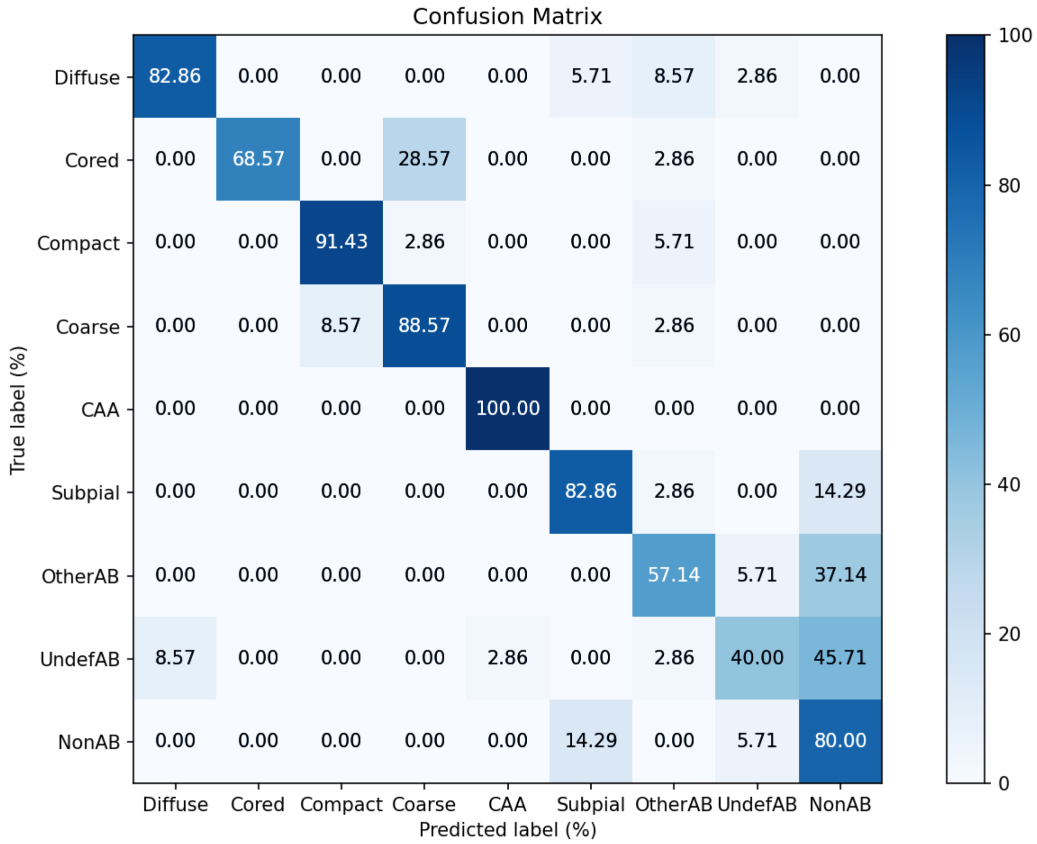


Figure 3.2: A confusion matrix depicting the overall classification performance of the five cross-validation models for all considered Aβ types. Rows in the matrix represent the true class labels, while the columns represent the predicted classifications. In other words, each cell in the matrix presents the percentage of instances predicted by the model versus the true class labels. Values are computed by aggregating the predictions across the 5-fold cross-validation test sets for each of respective the models.

Performance on Primary Aβ Types

After introducing the anomaly classifications, the diffuse class experienced its largest change in performance for the precision metric, decreasing from 100% to 92.14%. This reduction harms the models' reliability for diffuse predictions. On the other hand, the misclassification of diffuse plaques as subpial deposits was reduced from 14.29% to 5.71%. This improvement not only enhances the precision of subpial predictions but shows that the anomaly types can act as filters for incorrect classifications in the primary types.

In contrast to the diffuse plaques, cored plaques maintain a perfect precision of 100%. However, cored recall is lowered from 71.43% to 68.57%, and the percentage of cored misclassified as coarse plaques remains unchanged at 28.57%. As a result, the anomaly types are in this case filtering correct cored predictions instead of the 28.57% incorrectly classified as coarse.

Compact plaques exhibit relatively stable performance, with all metrics remaining above 90%. While the models managed to filter 2.85% of compact plaques incorrectly classified as coarse to the OtherAB anomaly type, 2.86% of correct predictions were also moved to the anomaly category. As a result, recall was reduced from 94.29% to 91.43%.

Coarse plaques exhibit an improvement in precision, with an increase from 72.89% to 74.00%. This enhancement results from reassigning compact plaques, which were previously misclassified as coarse, to the OtherAB anomaly class. The models effectively function as a filter in this case. On the other hand, the models perform somewhat undesirably in terms of recall of coarse plaques, which decreased from 91.43% to 88.57% with 2.86% of the plaques misclassified as OtherAB anomaly type. Furthermore, incorrect classifications of coarse as compact plaques remains unchanged at 8.57%.

CAA maintains a value of 100% in recall, while its precision and F1 values are lowered to 97.5% and 98.67%, respectively. Overall, including the anomaly types seems to have relatively minor effects on the average performances of CAA (highest metric difference is 2.5%).

Subpial deposits experienced a decline in performance for recall, from 100% to 82.86%. Figure 3.2 indicates the reason for this as 17.14% of subpial is misclassified as an anomaly plaque, in part revealing failure of the models in filtering only anomaly or incorrect classifications.

Performance on Anomaly Aβ Types

OtherAB shows a lower performance than at least five out of six primary classes for all its metrics, as depicted in Table 3.2. Although its lowest values are in recall at 57.14%, the remaining 42.86% of OtherAB misclassifications are assigned to other anomaly categories. Consequently, the models manage to at least correctly assign all OtherAB deposits as anomalies, fulfilling its role as an anomaly filter in this regard. However, the confusion matrix in Figure 3.2 also shows undesirable behavior where five out of the six primary classes were partly incorrectly classified as OtherAB, with the highest value at 8.57% for diffuse plaques. Additionally, it shows greatest instability for precision (74.22%) with a standard deviation of 20.46%, resulting in this measure being less reliable.

While UndefAB has a higher precision at 83.33% than OtherAB (74.22%), it has the lowest values among all anomalies for recall (40.00%), and F1 score (51.56%). Additionally, 11.43% of UndefAB are misclassified as primary classes. Its precision also shows the highest standard deviation among all types at 23.57%, revealing that its performance in that regards is less reliable. However, an advantage in UndefAB is that its false positives are mainly constrained to the anomaly categories, with the only primary false positive being 2.86% of diffuse plaques. As a result, the models avoid incorrectly capturing 5/6 of the primary types as anomaly UndefAB deposits.

NonAB has the highest recall at 80.0% with the lowest precision at 48.06% among the anomalies. While NonAB has the lowest precision, most of the false positives belong to the anomaly types, with the only primary false positive being 14.29% of subpial deposits. Ultimately, if the false positives in precision for one anomaly class mainly come from another anomaly, it is trivial on the models' performance in predicting the six primary classes.

3.1.3. Visualisations of Plaque Classifications

The ensemble model's predictions for a randomly chosen subset of plaques from the annotated dataset are shown in Figure 3.3. On this sample of data, the model displays mostly correct predictions, as indicated by the lack of blue text annotations in 83 out of the 90 depicted plaques. Diffuse, coarse, and subpial primary classes each have one false positive plaque respectively belonging to UndefAB, cored, and diffuse classes. Similarly, the OtherAB class has three false positives, where one originally belongs to compact and two to cored plaques. On the other hand, the classification visualization has no misclassifications for cored, compact, CAA, and UndefAB types on this sample of labeled data.

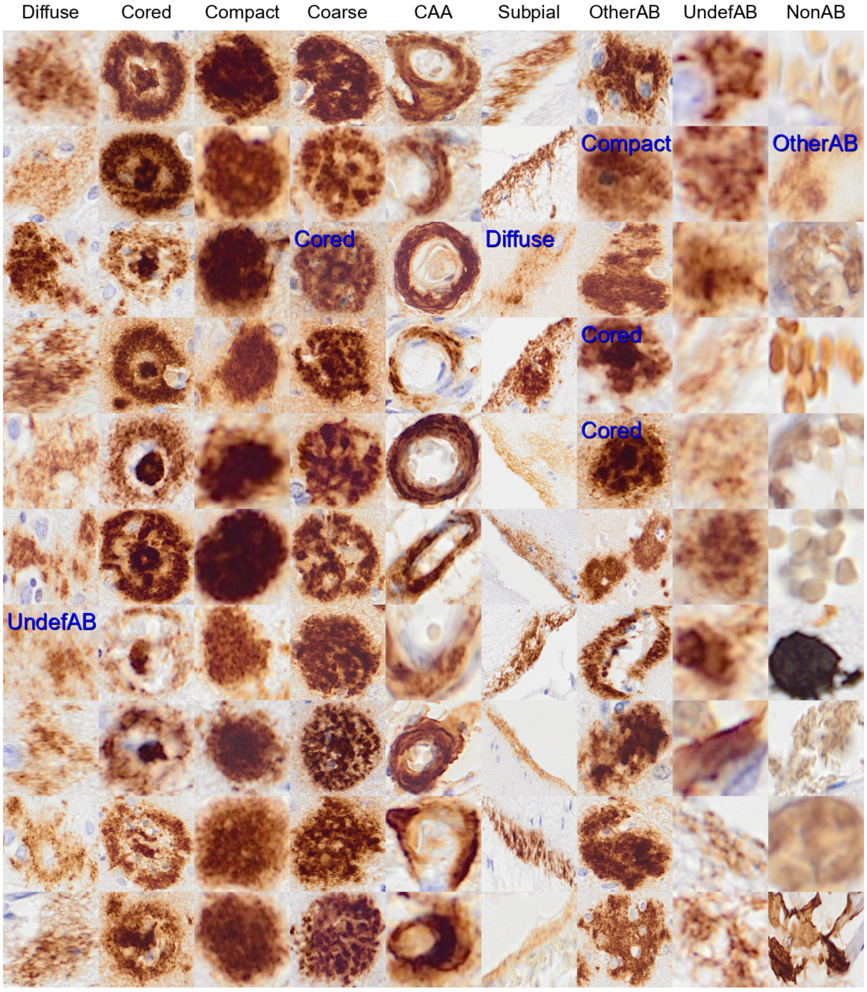


Figure 3.3: An image matrix composed of random samples from the labeled dataset (315 plaques in total). Each column represents a specific Aβ class, with each cell containing a random plaque image that the ensemble model has predicted for the class. Incorrectly classified plaques are annotated with blue text indicating their true class. Plaques without blue texts in this matrix are correctly classified by the model.

The predictions for randomly selected plaques from the unlabeled dataset are depicted in Figure 3.4. Overall, the image matrix seems to show believable predictions, mainly evident in the diffuse, cored, compact, CAA plaques. Although, some cored plaques with a hollow inner circle seem to be misclassified. For the NonAB category, it can be seen that its classifications contain plaques which are indeed not Aβ. Overall, discernible patterns can be distinguished from the classifications for this random unlabeled sample.

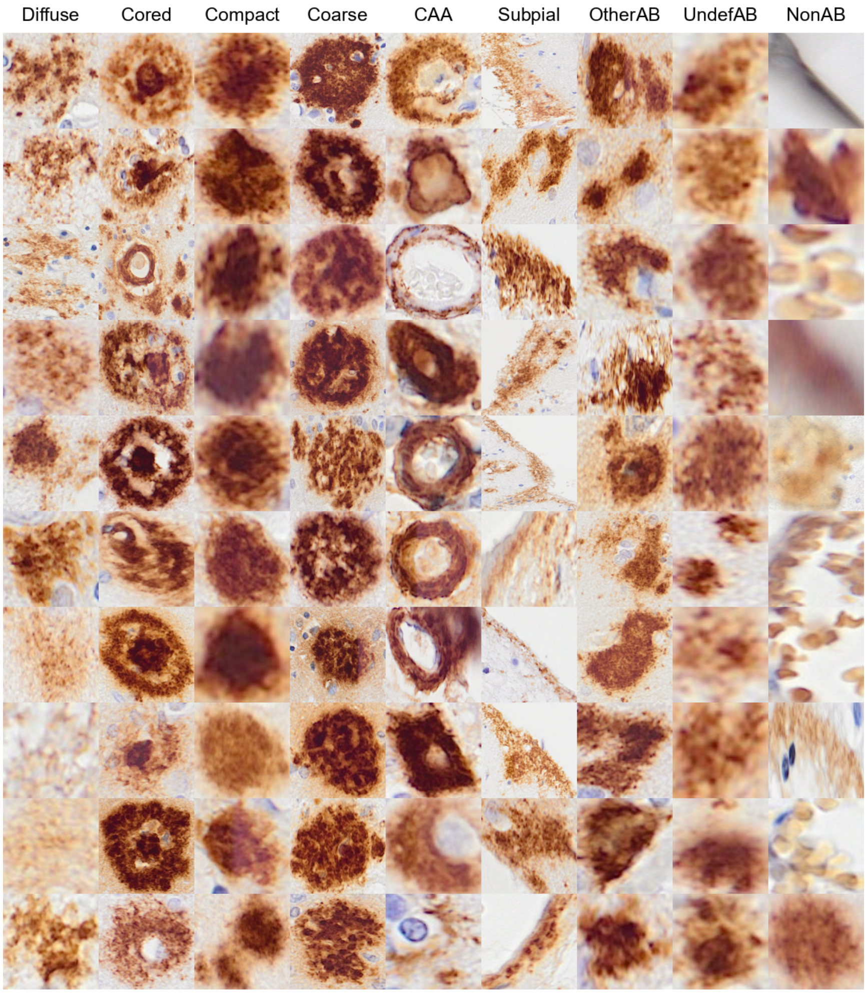


Figure 3.4: An image matrix with random samples from the unlabeled dataset, which contains around four million plaques in total. Each column represents a specific Aβ class, with each cell containing an example plaque image that the ensemble model has classified into that class.

3.2. Model Application Analysis

3.2.1. Distribution of Predicted Aβ Types

Plaque Frequency Distribution

The frequency distribution of all Aβ types predicted by the ensemble model on the unlabeled Aβ dataset is examined. Figure 3.5 shows that the majority of classified plaques are assigned to the anomaly categories, accounting for 60.86% of the 100+ and 60.78% of the AD cohort predictions. Among these anomalies, the UndefAB category has the highest proportion, with 43.84% of 100+ and 41.06% of AD.

The focus is changed to a frequency distribution that only includes plaques predicted to be a primary Aβ type. Figure 3.6 shows that diffuse plaques constitute the largest share of the predicted primary types, with 64.89% of 100+ and 70.73% of AD. The second most common category is compact plaques, representing 20.66% of 100+ and 18.03% of AD. CAA and subpial plaques have the lowest frequencies, with CAA at 3.04% of 100+ and 1.91% of AD, and subpial at 1.73% of 100+ and 1.06% of AD.

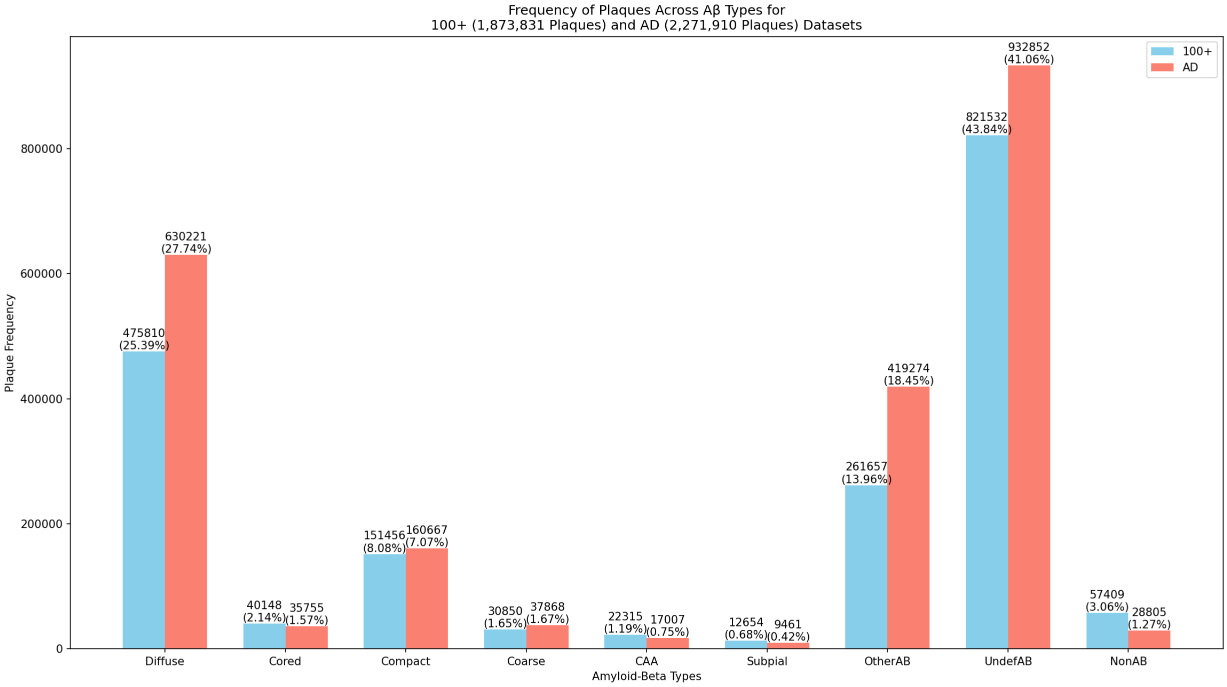


Figure 3.5: Frequency distribution of primary and anomaly Aβ types predicted by the model for the 100+ and AD cohorts. The entire unlabeled Aβ dataset is considered. The y-axis displays the number of plaques assigned to the class, while the x-axis the corresponding class. Values are given at the top of each bar, which reflect the absolute and relative frequency calculated for each cohort.

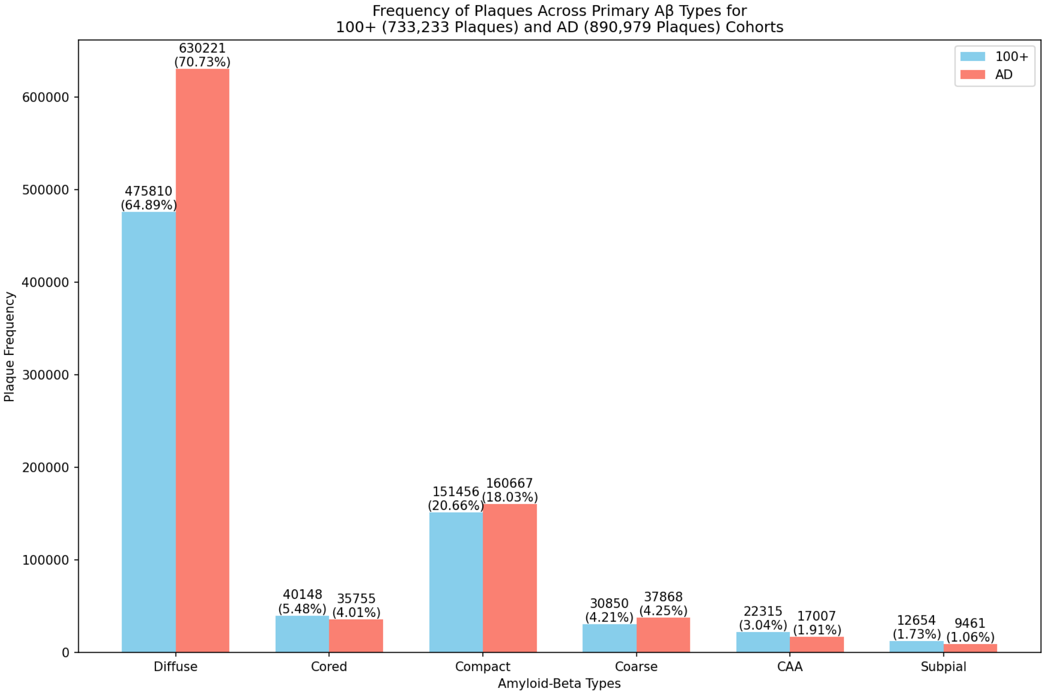


Figure 3.6: Frequency distribution of primary Aβ types predicted by the model for the 100+ and AD cohorts. Only the predicted primary types are considered for the unlabeled Aβ dataset. The y-axis displays the number of plaques assigned to the class, while the x-axis the corresponding class. Values are given at the top of each bar, which reflect the absolute and relative frequency calculated for each cohort on only the primary type predictions.

General Primary Aβ Load Distribution

Differences in Aβ load distribution between the 100+ and AD cohorts are analysed for each class. The Mann-Whitney U test reveals that these differences are statistically significant between cohorts in each class with p-values < 0.001, as indicated by the triple asterisks in Figure 3.7.

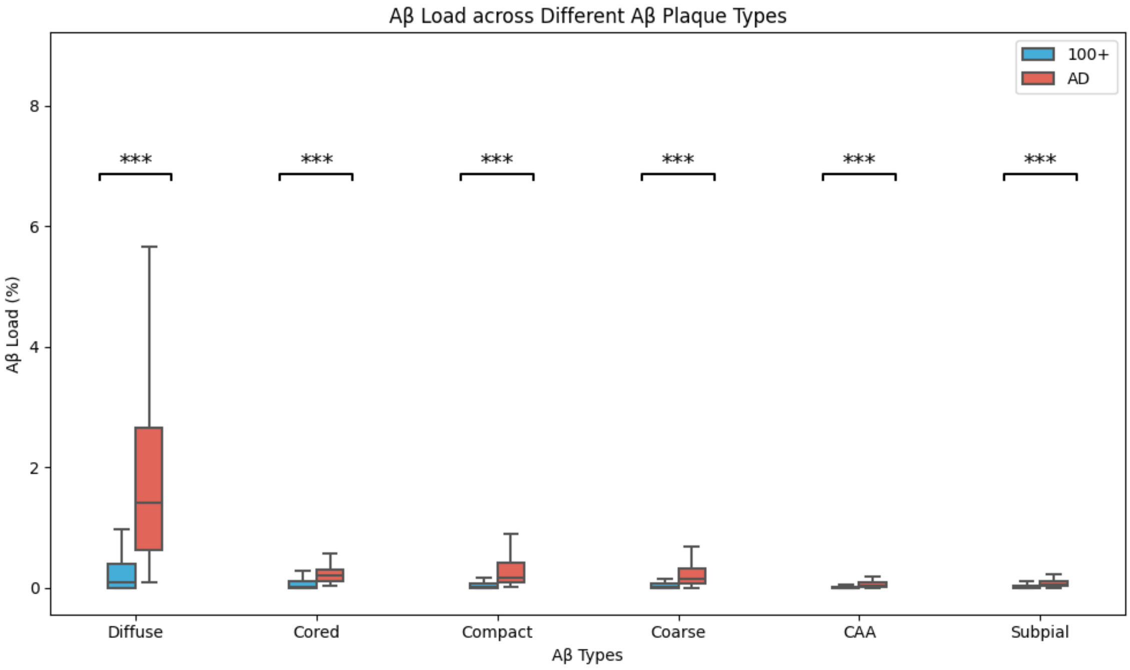


Figure 3.7: Box plot of the Aβ load distribution for the Aβ types predicted by the ensemble model. Each type's load distribution is given for the 100+ (in blue) and AD (in red) datasets. Loads are given as the percentage of grey matter that is covered in Aβ deposits of a class. Statistical significance is assessed between the cohorts for each class using the Mann-Whitney U test. Asterisks above the box plots indicate p-value significance levels: $* = p < 0.05$, $** = p < 0.01$, $** = p < 0.001$. Each box in the plot represents the IQR, with the median value indicated by the line inside the box. The whiskers that extend from the box show the range of the data within 1.5 times the IQR.

The diffuse class shows the largest difference between predicted cohort medians, with the 100+ median (0.0948% Aβ load) being lower than the AD median (1.4079% Aβ load). It also has the highest interquartile range (IQR), with lower values for 100+ (Q1: 0.0053%, Q3: 0.3993% Aβ load) compared to AD (Q1: 0.6353%, Q3: 2.6589% Aβ load). The higher Aβ load for diffuse plaques compared to other primary types coincides with the findings of higher frequency in Figure 3.6.

Cored, compact, coarse, and subpial classes support the idea that Aβ load range and medians are lower for 100+ than AD. For the cored class, the AD median (0.2114% Aβ load) surpasses the 100+ median (0.0217% Aβ load). Its IQR is also lower for 100+ (Q1: 0.0%, Q3: 0.1179% Aβ load) in contrast to AD (Q1: 0.1105%, Q3: 0.3107% Aβ load). Similarly, in the case of compact plaques, the 100+ median (0.0071% Aβ load) is smaller compared to the AD median (0.1673% Aβ load). Compact's IQR for 100+ (Q1: 0.0002%, Q3: 0.0705% Aβ load) is also smaller than AD (Q1: 0.0862%, Q3: 0.4084% Aβ load). Likewise, the coarse class shows a 100+ median (0.0112% Aβ load) lower than the AD median (0.1408% Aβ load). Additionally, its IQR for 100+ (Q1: 0.0%, Q3: 0.0642% Aβ load) is lower as well than for AD (Q1: 0.0716%, Q3: 0.3185% Aβ load). Subpial's 100+ median (0.0124% Aβ load) also falls short compared to AD (0.0622% Aβ load). Its IQR is smaller for 100+ (Q1: 0.0008%, Q3: 0.0425% Aβ load) as well against AD (Q1: 0.0384%, Q3: 0.1144% Aβ load).

Compared to the values of the other primary Aβ types loads, CAA shows the lowest median Aβ load difference of 0.0297% between cohorts, with the 100+ median (0.0051% Aβ load) once more being lower than the AD median (0.0348% Aβ load). Furthermore, its IQR is smaller for 100+ (Q1: 0.0001%, Q3: 0.0191% Aβ load) than for AD (Q1: 0.0140%, Q3: 0.0919% Aβ load).

Overall, the general trend for every primary Aβ type is that the 100+ cohort has lower Aβ loads compared to the AD cohort. Likewise, the Aβ load IQR is also smaller for the 100+ cohort for all types. In addition, diffuse plaques reveal a higher median Aβ load difference between the 100+ and AD cohorts, while CAA a lower one.

Primary Aβ Load Distribution Across Cerebral Regions

To understand how Aβ loads vary between 100+ and AD cohorts across different cerebral regions, an overview of regional loads for each Aβ type predicted by the ensemble model is given in Figure 3.8. The plotted subfigures reveal that differences between cohorts within the regional loads are statistically significant for every class with p-value < 0.001.

For the diffuse plaques, the medians for Aβ load in the frontal, parietal, temporal, and occipital regions are lower for 100+ (0.1714%, 0.0981%, 0.0607%, 0.0480% Aβ load, respectively) than that of the AD cohort (2.7052%, 1.8423%, 1.0271%, 0.6422%, Aβ load, respectively). Furthermore, the region with the higher difference (2.5338% Aβ load) between cohort medians is the frontal region, while the lower difference (0.5942% Aβ load) belongs to the occipital region.

In the case of the cored plaques, the median Aβ loads in the frontal, parietal, temporal, and occipital regions are also lower for the 100+ cohort (0.0444%, 0.0310%, 0.0421%, 0.0104% Aβ load, respectively) compared to the AD cohort (0.2189%, 0.2448%, 0.1159%, 0.2303% Aβ load, respectively). Additionally, the occipital region exhibits the highest difference in medians (0.2199% Aβ load), whereas the temporal region the lowest difference (0.0738% Aβ load).

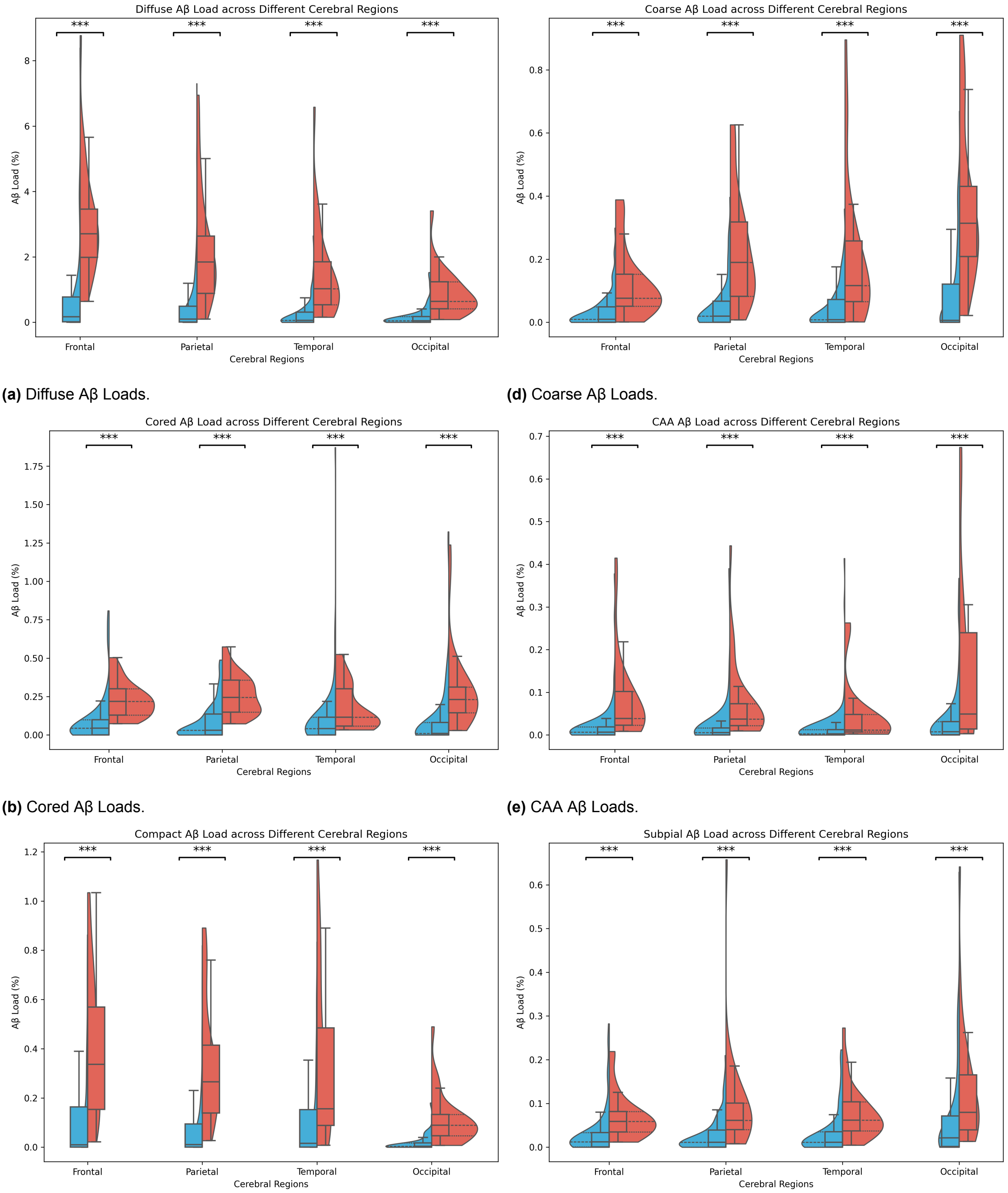
Regarding the compact plaques, the median Aβ loads in the frontal, parietal, temporal, and occipital regions are smaller for the 100+ cohort (0.0097%, 0.0109%, 0.0155%, 0.0041% Aβ load, respectively) in relation to the AD cohort (0.3369%, 0.2659%, 0.1562%, 0.0895% Aβ load, respectively). The frontal region displays the greatest difference in medians (0.3272% Aβ load), while the occipital region the smallest difference (0.0854% Aβ load).

When examining the coarse plaques, the median Aβ loads in the frontal, parietal, temporal, and occipital regions are less for the 100+ cohort (0.0096%, 0.0193%, 0.0086%, 0.0065% Aβ load, respectively) in comparison to the AD cohort (0.0762%, 0.1898%, 0.1161%, 0.3140% Aβ load, respectively). In addition, the occipital region shows the highest difference in medians (0.3075% Aβ load), while the frontal region shows the lowest difference (0.0666% Aβ load).

For CAA, the median Aβ loads in the frontal, parietal, temporal, and occipital regions are lower for the 100+ cohort (0.0066%, 0.0056, %, 0.0026%, 0.0078% Aβ load, respectively) than for the AD cohort (0.0386%, 0.0374%, 0.0114%, 0.0492% Aβ load, respectively). The occipital region shows the highest difference in medians (0.0414% Aβ load), whereas the temporal region shows the lowest difference (0.0088% Aβ load).

Considering the subpial deposits, the median Aβ loads in the frontal, parietal, temporal, and occipital regions are reduced for the 100+ cohort (0.0124%, 0.0111%, 0.0111%, 0.0213% Aβ load, respectively) compared to the AD cohort (0.0588%, 0.0615%, 0.0617%, 0.0794% Aβ load, respectively). Furthermore, the occipital region exhibits the highest difference in medians (0.0581% Aβ load), while the frontal region the lowest (0.0464% Aβ load).

Overall, across all plaque types and cerebral regions, the comparisons reveal that the 100+ cohort consistently has lower Aβ loads compared to the AD cohort.



(a) Diffuse Aβ Loads.

(d) Coarse Aβ Loads.

(b) Cored Aβ Loads.

(e) CAA Aβ Loads.

(c) Compact Aβ Loads.

(f) Subpial Aβ Loads.

Figure 3.8: Violin box plots of Aβ load distributions displayed per Aβ type. Each plot displays loads for the 100+ (in blue) and AD (in red) datasets over the frontal, parietal, temporal, and occipital cerebral regions. Loads were calculated from predictions by the ensemble model. (a) Shows distribution for diffuse Aβ loads. (b) Shows distribution for cored Aβ loads.(c) Shows distribution for compact Aβ loads. (d) Shows distribution for coarse Aβ loads. (e) Shows distribution for CAA Aβ loads. (f) Shows distribution for subpial Aβ loads. Statistical significance is assessed between the cohorts for each region using the Mann-Whitney U test. Asterisks above the box plots indicate p-value significance levels: $* = p < 0.05$, $** = p < 0.01$, $** = p < 0.001$.

3.2.2. Correlations Between Predicted Aβ Loads and 100+ Assessments

Correlations for Neuropathological Staging Schemes

The correlations between the Aβ type loads and different neuropathological staging schemes are examined. First, Figure 3.9a shows that only statistically significant positive correlations are present in the confusion matrix (p-values < 0.001). Furthermore, the correlations reveal fair to moderate behavior across the different Aβ types, with the correlation coefficient r ranging from 0.46 to 0.73. For the Thal Aβ phase, a fair correlation is shown for subpial ($r = 0.59$), while the other classes show moderate correlations (r ranging 0.65-0.73). In the case of the Thal CAA stage, the strongest correlation is observed for CAA ($r = 0.66$). Furthermore, the CERAD NP scores show its highest correlation for cored plaques ($r = 0.68$).

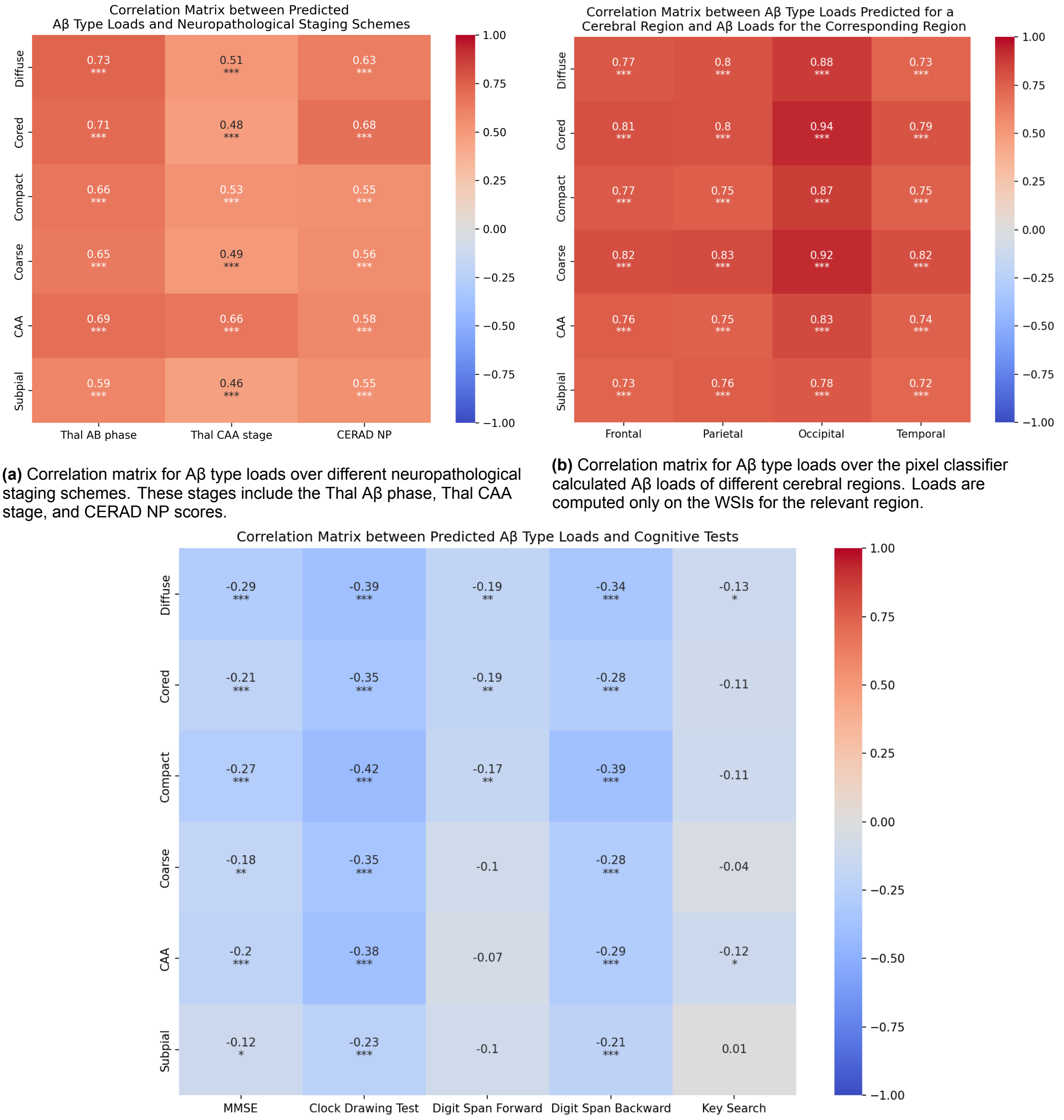
Since Thal CAA and CERAD NP scores are expected to somewhat measure and show strongest correlations with CAA and cored, these variables are considered for further analysis. Figure 3.10 illustrates the relationship between the model's predicted CAA Aβ loads and Thal CAA stages, presented as a scatter plot overlaid with box plots. The median values of the box plots progressively increase with each stage (median Aβ load across stages: 0.0%, 0.007%, 0.048%, 0.292%), with statistically significant differences between the stages (p-values < 0.001). For CERAD NP, Figure 3.11 displays the predicted Aβ load for cored plaques in relation to CERAD NP scores. The median values generally increase until CERAD NP score 2 (median Aβ load across first three stages: 0.0%, 0.063%, 0.100%). There is a decrease in median for the last score 3 (0.066% Aβ load), but it is not statistically significant compared to the distribution of score 2.

Correlations for Aβ Loads of Cerebral Regions

An analysis was conducted on the correlations between Aβ loads pre-calculated by a pixel level classifier for a specific cerebral region and the loads predicted by the ensemble model for Aβ types from WSIs of that region. All relationships displayed in Figure 3.9b show moderate to very strong positive correlations that are statistically significant, with r ranging from 0.72 to 0.94. Positive correlations are expected since an increase in general Aβ load for a region should lead to an increase in Aβ load for at least one the types. Overall, the strongest correlations are found across all classes for the Aβ loads in the occipital region (r ranging 0.78-0.94). Cored (r ranging 0.79-0.94) and coarse (r ranging 0.82-0.92) classes show the highest correlations among the types over all cerebral regions. Additionally, cored ($r = 0.94$) and coarse ($r = 0.92$) loads exhibit stronger correlations for the occipital region compared to other regions. On the other hand, compared to other classes subpial displays the lowest values (r ranging 0.72-0.78) for all of the regions except for parietal.

Correlations for Cognitive Tests

Several negative relationships are observed between Aβ loads for different classes and scores from cognitive assessments completed by the centenarians. Diffuse loads in Figure 3.9c show poor correlations with MMSE ($r = -0.29, p < 0.001$), DSF ($r = -0.19, p < 0.01$), and KS ($r = -0.13, p < 0.05$). On the other hand, they show fair correlations with CDT ($r = -0.39, p < 0.001$) and DSB ($r = -0.34, p < 0.001$). Similarly, the cored class demonstrates a fair correlation with CDT ($r = -0.35, p < 0.001$). However, it shows poor correlations with other statistically significant tests: MMSE ($r = -0.21, p < 0.001$), DSF ($r = -0.19, p < 0.01$), and DSB ($r = -0.28, p < 0.001$). The compact class behaves similarly to diffuse plaques, with fair negative correlations for CDT ($r = -0.42, p < 0.001$) and DSB ($r = -0.39, p < 0.001$), while other statistically significant tests show poor correlations (r ranging from -0.17 to -0.27). The coarse class only has statistically significant correlations with MMSE ($r = -0.18$), CDT ($r = -0.35$), and DSB ($r = -0.28$). For CAA, similar patterns are observed with MMSE ($r = -0.2, p < 0.001$), CDT ($r = -0.38, p < 0.001$), DSB ($r = -0.29, p < 0.001$), and KS ($r = -0.12, p < 0.05$). Subpial shows statistically significant poor negative correlations with MMSE ($r = -0.12$), CDT ($r = -0.23$), and DSB ($r = -0.21$). Overall, class types exhibit the strongest correlations with CDT and DSB assessments.



(a) Correlation matrix for Aβ type loads over different neuropathological staging schemes. These stages include the Thal Aβ phase, Thal CAA stage, and CERAD NP scores.

(b) Correlation matrix for Aβ type loads over the pixel classifier calculated Aβ loads of different cerebral regions. Loads are computed only on the WSIs for the relevant region.

(c) Correlation matrix for Aβ type loads over scores of different cognition assessment tests.

Figure 3.9: Correlation matrices for the predicted Aβ loads of different plaque types over the assessment data included alongside the 100+ WSIs. Correlation are computed through Spearman [32]. The degree of statistical significance is denoted by asterisks: $* = p < 0.05$, $** = p < 0.01$, $*** = p < 0.001$. (a) Correlation matrix for Aβ type loads over different neuropathological stages. (b) Correlation matrix for Aβ type loads over the pixel classifier Aβ loads of different cerebral regions. (c) Correlation matrix for Aβ type loads over cognition test scores.

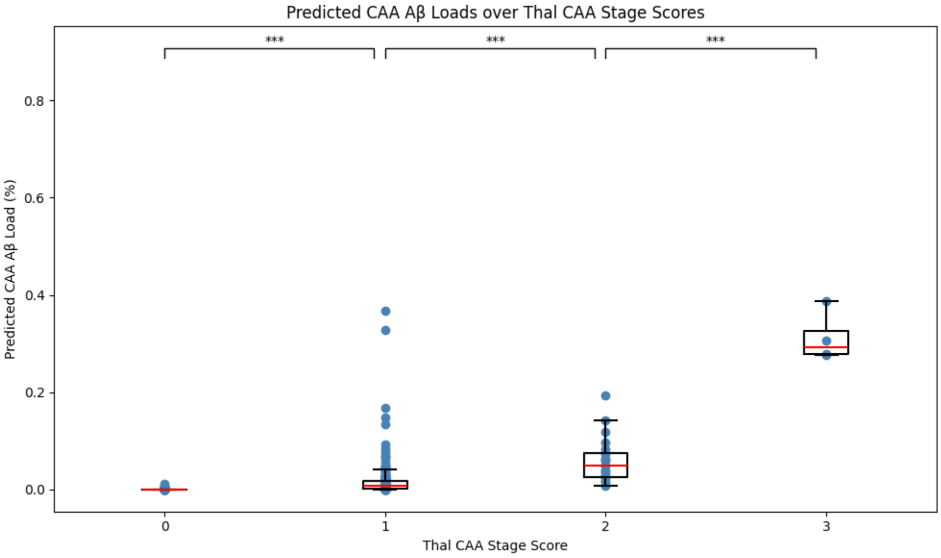


Figure 3.10: Scatter plot of the predicted CAA Aβ loads against the Thal CAA stages assigned to the corresponding centenarians. Stages range from zero to three. Box plots are plotted for each stage based on the calculated CAA Aβ loads.

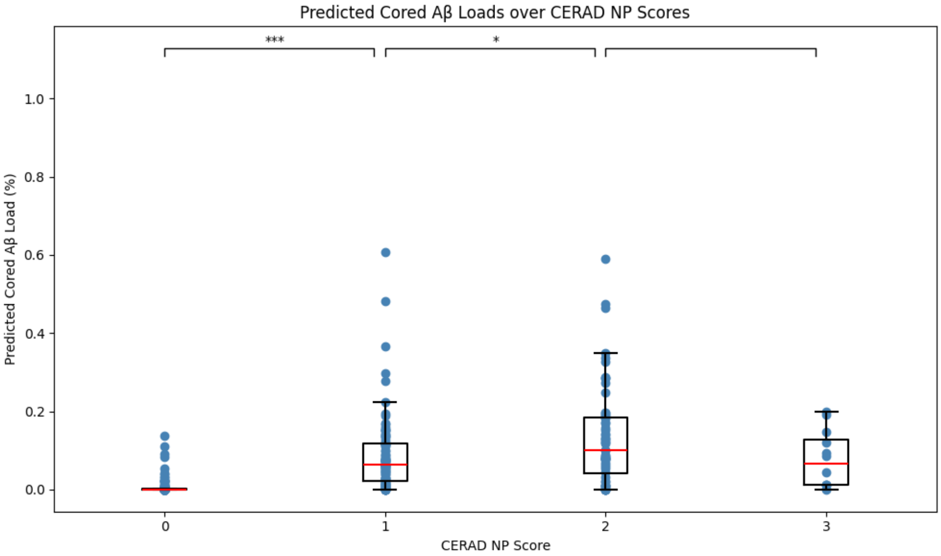


Figure 3.11: Scatter plot of the predicted cored Aβ loads against the CERAD NP scores assigned to the corresponding centenarians. Stages range from zero to three. Box plots are plotted for each stage based on the calculated cored Aβ loads

4

Discussion

4.1. Model Performance

The performance estimates for primary and anomaly Aβ types, detailed in Section 3.1.2, indicate that the fine-tuned ResNet50 model is an effective classifier for the primary classes, with an average accuracy of 85.71% and precision of 89.47%. For individual primary types, cored plaques exhibit the lowest performance in terms of recall (68.57%). However, they compensate with a perfect precision (100%), which is beneficial for ensuring that the Aβ class predictions of the model can be relied on to actually contain the class. Conversely, the coarse class has the lowest precision (74.0%), leading to less reliable predictions where coarse plaques may include incorrect classifications of other types. This could negatively impact the reliability of Aβ distribution analyses for coarse plaques in the final model. In contrast, the model achieves perfect recall and high precision (97.5%) for CAA, indicating that predictions for this class should be reliable.

While the fine-tuned models' performance on anomaly types (59.05% recall, 68.54% precision) is revealed to be weaker than that of the primary types, this is acceptable since anomalies are not the primary focus of the analysis and instead serve as a filter that can prune irregular Aβ deposits from the primary types. In that regards, the model has been effective in anomaly detection, as evinced by the confusion matrix in Section 3.1.2. The confusion matrix shows that, on average, 91.43% of the ground truth anomalies are contained within the three anomaly types, while only an average of 6.67% of the primary plaques are misclassified as anomalies. Overall, the model seems to fulfill its secondary role as an anomaly filter.

The visualizations of classifications in Section 3.1.3 reveal discernible patterns for the primary Aβ types on the unlabeled Aβ dataset. However, a few misclassifications do occur on the unseen dataset, particularly for cored plaques, which are not fully represented in the confusion matrix that shows 100% precision for cored on the test sets. This issue is likely due to the limited data available, a limitation posed by the few-shot learning problem, which restricts dataset diversity and sample representation of the original dataset. To more reliably assess the final performance of the ensemble model, it is recommended to use a separate annotated holdout test set, which is not involved in the cross-validation process.

4.2. Model Application

The Aβ load distributions in Section 3.2.1 indicate that, for each primary Aβ type, the Aβ loads are significantly lower in the 100+ cohort compared to the AD cohort. This relationship also holds across all cerebral regions for each class. These results align with existing findings that state that Aβ load in centenarians is significantly lower compared to AD individuals for the same brain regions [11].

The correlations depicted in Section 3.2.2 partially validate the model's performance. Thal Aβ phases exhibit statistically significant positive fair to moderate correlations (r ranging from 0.59 to 0.73) across all plaque types. This agrees with existing research, as higher Thal Aβ measures greater Aβ spread across the brain [12]. Additionally, the Thal CAA stage exhibits a strong positive correlation with CAA ($r = 0.66, p < 0.001$), aligning with literature since the Thal CAA stage represents the severity of CAA [13]. This is further supported by the scatter plot in Figure 3.10, which demonstrates that the median Aβ loads significantly increase with the progression of CAA stages. Similarly, CERAD NP scores, which measure neuritic plaques (a subtype of cored plaques) [14], exhibit a higher correlation with cored plaques ($r = 0.68, p < 0.001$). This is corroborated by the scatter plot in Figure 3.11, which illustrates that median Aβ loads increase with CERAD NP scores between populations that are statistically significant. These findings seem to indicate proper predictions by the model.

Considering that higher scores in the cognitive tests completed by the centenarians indicate better cognition, some negative correlations with Aβ types were expected. Likewise, Section 3.2.2 reveals that all Aβ types show some degree of statistically significant negative correlations to the assessments, with r ranging from -0.17 to -0.42. These correlations, for each Aβ type, are more pronounced for the CDT (r ranging -0.23 to -0.42) and DSB (r ranging -0.21 to -0.39) assessments, which is consistent with other studies showing stronger negative correlations between these tests and general Aβ loads [11]. The compact plaque, which is positively associated with the severity of AD [15], also shows stronger significant negative correlations with the CDT ($r = -0.42$) and DSB ($r = -0.39$) tests compared to other types.

4.3. Future Work and Limitations

The primary limitation of this work is the size of the annotated dataset. For future work, it is recommended to evaluate the final ensemble model's performance on a separate test set. Currently, the performance estimate is based on the average results from five different models. This may differ from the actual performance of the final model where predictions are obtained through majority voting. Additionally, some of the standard deviations (>5%) for the models' performance estimates suggest that the model may be overfitting on the training data. Increasing the amount of training data for unstable classes can help mitigate overfitting. Otherwise, further exploration in changing the parameters for regularisation techniques such as data augmentation, dropout probability, and weight decay could be conducted. Additionally, k in k-fold cross-validation can be assessed thoroughly to determine a more effective combination.

Other future work for the model can involve using contrastive learning techniques to leverage the large amounts of unlabeled Aβ data. In the initial stages of the research, some unsupervised contrastive learning models were explored [25], [26], before opting for a ResNet50 encoder pre-trained on the ImageNet dataset. Unfortunately, pre-training a contrastive learning model on the unlabeled Aβ dataset was not found to yield improvements for the supervised fine-tuned encoder. However, there are other unexplored methods that integrate annotated data in the contrastive learning process, such as the Supervised Momentum Contrastive learning (SupMoCo) model [34]. This approach can be adapted to work with the few samples available for few-shot learning, potentially capturing more suitable representations.

Another improvement for the fine-tuned model can be achieved by refining its threshold selection method. Currently, thresholds are chosen to balance a high true positive rate and a low false positive rate. However, this approach may not be as suitable for analysing classified Aβ plaques. For example, the threshold selection could be adjusted to prioritize a lower false positive rate for the primary plaque types, ensuring that predictions more reliably belong to those classes. It is recommended to carefully adjust the thresholds with feedback from domain experts to determine to what extent the threshold should prioritize one measure over the other.

As another direction, Gradient-weighted Class Activation Mapping (GradCAM) could be used to enhance the model's explainability [35]. GradCAM can visualize areas within plaque images that the model relies on for its predictions. In this manner, insights could be gained into whether the model focuses on the appropriate features or if it is overfitting to irrelevant ones.

For other future work, a final extension to the Aβ classification pipeline can be incorporated. This could include an interface that facilitates the inspection of predicted plaques and enables an easier process for adding new annotations.

5 Conclusion

Given the negative effect of AD on cognitive health in the aging population, it would be beneficial to gain a better understanding of the disease. A characteristic in the pathology of AD are Aβ plaques found in grey matter, where these deposits can appear in various forms. The six primary plaques of interest in this study are: diffuse plaques, cored plaques, compact plaques, coarse-grained plaques, CAA, and subpial aggregation. Differences between healthy aging and AD can be analysed by comparing these Aβ types between centenarians in the 100+ study and individuals with AD. However, due to the large number of plaques per WSI, manual annotation of the Aβ types becomes costly. To address this challenge, an Aβ detection pipeline was used to segment grey matter from the brain tissue WSIs, locate plaques within the grey matter, and classify the plaques. This research covered the classification stage, where a ResNet50 model was fine-tuned on a limited number of annotations (315) for the classification of the primary Aβ types in the WSIs. To handle unknown or irregular plaques in the dataset, the model's training also included annotations for three anomaly Aβ types whose purpose was to filter out unconventional plaques in the unlabeled Aβ data from the primary ones.

The ResNet50 model was originally pre-trained on the ImageNet dataset through contrastive learning by MoCo [26]. Following this, the ResNet50 encoder was extracted and fine-tuned using annotations from the six primary and three anomaly types of Aβ plaques. For each of the nine types, only 35 annotations were used. Five ResNet50 models were fine-tuned with 5-fold cross-validation on different dataset configurations to reduce prediction instability from only using a single model. By averaging estimates across the test set folds, the models demonstrated an average performance of 85.71% accuracy and 89.47% precision for classifying the primary types. An ensemble model, composed of all five trained models, was used for final classifications of Aβ plaques by majority voting.

For an analysis comparing Aβ distributions between the 100+ and AD cohorts, Aβ loads were calculated based on the primary Aβ types predicted by the ensemble model. Results revealed that, across all Aβ types, the Aβ loads in the 100+ cohort were significantly lower than those in the AD cohort. This pattern persisted across the frontal, parietal, temporal, and occipital regions for each Aβ type. These findings aligned with existing literature indicating that Aβ loads were found to be significantly lower in centenarians compared to individuals with AD [11].

To somewhat validate the performance of the model on the unlabeled Aβ dataset, benchmarks in neuropathological assessment were used, which include the Thal Aβ phase [12], Thal CAA stage [13], and CERAD NP scores [14]. The Thal Aβ phase measures the spread of Aβ in the brain [12] and showed positive correlations to all predicted Aβ load types (r ranging 0.59-0.73, p > 0.001). The Thal CAA stage evaluates the severity of CAA and had a stronger positive correlation with CAA ($r = 0.66, p > 0.001$) among the primary Aβ types. Additionally, the CERAD NP score assesses the density of neuritic plaques, which are subset of cored plaques, and had a stronger positive correlation to cored plaques ($r = 0.68, p > 0.001$). Overall, predicted results from the model seemed to align with standard measurements used in literature, suggesting suitable performance on the unlabeled Aβ dataset.

For future work, the performance of the model may improve if pre-training is completed on the unlabeled

Aβ dataset instead of ImageNet, using semi-supervised contrastive learning methods [34]. Additionally, to be able to reliably assess the final performance of the ensemble model, it is recommended to use an annotated test set separate from the cross-validation data. Finally, to finish the Aβ detection pipeline, an interface could be developed to examine the classified Aβ plaques.

References

[1] A. S. Schachter and K. L. Davis, “Alzheimer’s disease,” *Dialogues in Clinical Neuroscience*, vol. 2, no. 2, pp. 91–100, Jun. 2000. DOI: `10.31887/dcns.2000.2.2/asschachter`.

[2] M. P. Murphy and H. LeVine, “Alzheimer’s disease and the amyloid-β peptide,” *Journal of Alzheimer’s Disease*, vol. 19, no. 1, pp. 311–323, Jan. 2010. DOI: `10.3233/jad-2010-1221`.

[3] H. M. Wisniewski, C. Bancher, M. Barcikowska, G. Y. Wen, and J. Currie, “Spectrum of morphological appearance of amyloid deposits in alzheimer’s disease,” *Acta Neuropathologica*, vol. 78, no. 4, pp. 337–347, 1989. DOI: `https://doi.org/10.1007/bf00688170`.

[4] H. Holstege, N. Beker, T. Dijkstra, *et al.*, “The 100-plus study of cognitively healthy centenarians: Rationale, design and cohort description,” *European Journal of Epidemiology*, vol. 33, no. 12, pp. 1229–1249, Dec. 2018. DOI: `https://doi.org/10.1007/s10654-018-0451-3`. [Online]. Available: `https://link.springer.com/article/10.1007%2Fs10654-018-0451-3`.

[5] A. Ganz, “The brains of cognitively healthy centenarians: What does the healthy aged brain tell us about alzheimer’s disease?” English, PhD-Thesis - Research and graduation internal, Vrije Universiteit Amsterdam, Dec. 2023. DOI: `10.5463/thesis.382`.

[6] N. Tesi, S. van der Lee, M. Hulsman, *et al.*, “Cognitively healthy centenarians are genetically protected against alzheimer’s disease,” *Alzheimer’s Dementia: The Journal of the Alzheimer’s Association*, Apr. 2024. DOI: `https://doi.org/10.1002/alz.13810`. [Online]. Available: `https://pubmed.ncbi.nlm.nih.gov/38634500/`.

[7] Z. Tang, K. V. Chuang, C. DeCarli, *et al.*, “Interpretable classification of alzheimer’s disease pathologies with a convolutional neural network pipeline,” *bioRxiv (Cold Spring Harbor Laboratory)*, Oct. 2018. DOI: `https://doi.org/10.1101/454793`.

[8] V. G. Ramaswamy, M. Ahirwar, G. Ryan, M. Signaevski, V. Haroutunian, and S. Finkbeiner, “A scalable high throughput fully automated pipeline for the quantification of amyloid pathology in alzheimer’s disease using deep learning algorithms,” *bioRxiv (Cold Spring Harbor Laboratory)*, May 2023. DOI: `https://doi.org/10.1101/2023.05.19.541376`.

[9] T.-L. Stephen, L. Korobkova, B. Breningstall, *et al.*, “Machine learning classification of alzheimer’s disease pathology reveals diffuse amyloid as a major predictor of cognitive impairment in human hippocampal subregions,” *bioRxiv (Cold Spring Harbor Laboratory)*, Jun. 2023. DOI: `https://doi.org/10.1101/2023.05.31.543117`.

[10] C. de Vries, “Amyloid-beta plaque quantification and analysis,” *repository.tudelft.nl*, 2023. [Online]. Available: `http://resolver.tudelft.nl/uuid:7b67aaa9-41a5-4bc3-b987-68f34ed145d9`.

[11] S. K. Rohde, P. Fierro-Hernández, A. J. Rozemuller, *et al.*, “Resistance to cortical amyloid-beta associates with cognitive health in centenarians,” *medRxiv*, 2023. DOI: `10.1101/2023.12.28.23300604`. eprint: `https://www.medrxiv.org/content/early/2023/12/29/2023.12.28.23300604.full.pdf`. [Online]. Available: `https://www.medrxiv.org/content/early/2023/12/29/2023.12.28.23300604`.

[12] D. R. Thal, U. Rüb, M. Orantes, and H. Braak, “Phases of aβ-deposition in the human brain and its relevance for the development of ad,” *Neurology*, vol. 58, no. 12, pp. 1791–1800, 2002. DOI: `10.1212/WNL.58.12.1791`. eprint: `https://www.neurology.org/doi/pdf/10.1212/WNL.58.12.1791`. [Online]. Available: `https://www.neurology.org/doi/abs/10.1212/WNL.58.12.1791`.

[13] D. R. Thal, E. Ghebremedhin, M. Orantes, and O. D. Wiestler, “Vascular Pathology in Alzheimer Disease: Correlation of Cerebral Amyloid Angiopathy and Arteriosclerosis/Lipohyalinosis with Cognitive Decline,” *Journal of Neuropathology Experimental Neurology*, vol. 62, no. 12, pp. 1287–1301, Dec. 2003, ISSN: 0022-3069. DOI: `10.1093/jnen/62.12.1287`. eprint: `https://academic.oup.com/jnen/article-pdf/62/12/1287/8133835/62-12-1287.pdf`. [Online]. Available: `https://doi.org/10.1093/jnen/62.12.1287`.

[14] S. S. Mirra, A. Heyman, D. McKeel, *et al.*, “The consortium to establish a registry for alzheimer’s disease (cerad),” *Neurology*, vol. 41, no. 4, pp. 479–479, 1991. DOI: 10.1212/WNL.41.4.479. eprint: https://www.neurology.org/doi/pdf/10.1212/WNL.41.4.479. [Online]. Available: https://www.neurology.org/doi/abs/10.1212/WNL.41.4.479.

[15] F. Liu, J. Sun, X. Wang, *et al.*, “Focal-type, but not diffuse-type, amyloid beta plaques are correlated with alzheimer’s neuropathology, cognitive dysfunction, and neuroinflammation in the human hippocampus,” *Neuroscience Bulletin*, vol. 38, Aug. 2022. DOI: 10.1007/s12264-022-00927-5.

[16] L. C. Walker, “Aβ plaques,” *Free Neuropathology*, vol. 1, pp. 1–31, Oct. 2020. DOI: https://doi.org/10.17879/freeneuropathology-2020-3025. [Online]. Available: https://www.uni-muenster.de/Ejournals/index.php/fnp/article/view/3025.

[17] B. Boon, M. Bulk, A. Jonker, *et al.*, “The coarse-grained plaque: A divergent aβ plaque-type in early-onset alzheimer’s disease,” English, *Acta Neuropathologica*, vol. 140, no. 6, pp. 811–830, Dec. 2020, ISSN: 0001-6322. DOI: https://doi.org/10.1007/s00401-020-02198-8.

[18] D. R. Thal, E. Ghebremedhin, U. Rüb, H. Yamaguchi, K. Del Tredici, and H. Braak, “Two Types of Sporadic Cerebral Amyloid Angiopathy,” *Journal of Neuropathology Experimental Neurology*, vol. 61, no. 3, pp. 282–293, Mar. 2002, ISSN: 0022-3069. DOI: 10.1093/jnen/61.3.282. eprint: https://academic.oup.com/jnen/article-pdf/61/3/282/8133497/61-3-282.pdf. [Online]. Available: https://doi.org/10.1093/jnen/61.3.282.

[19] T. Le, R. Crook, and J. Hardy, “Cotton wool plaques in non-familial late-onset alzheimer disease,” *Journal of neuropathology and experimental neurology*, vol. 60, pp. 1051–61, Dec. 2001.

[20] E. Richard, A. Carrano, J. Hoozemans, *et al.*, “Characteristics of dyshoric capillary cerebral amyloid angiopathy,” *Journal of neuropathology and experimental neurology*, vol. 69, pp. 1158–67, Oct. 2010. DOI: 10.1097/NEN.0b013e3181fab558.

[21] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” *CoRR*, vol. abs/1512.03385, 2015. arXiv: 1512.03385. [Online]. Available: http://arxiv.org/abs/1512.03385.

[22] M. B. Hossain, S. H. S. Iqbal, M. M. Islam, M. N. Akhtar, and I. H. Sarker, “Transfer learning with fine-tuned deep cnn resnet50 model for classifying covid-19 from chest x-ray images,” *Informatics in Medicine Unlocked*, vol. 30, p. 100 916, 2022, ISSN: 2352-9148. DOI: https://doi.org/10.1016/j.imu.2022.100916. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S235291482200065X.

[23] L. Zhang, Y. Bian, P. Jiang, and F. Zhang, “A transfer residual neural network based on resnet-50 for detection of steel surface defects,” *Applied Sciences*, vol. 13, no. 9, 2023, ISSN: 2076-3417. DOI: 10.3390/app13095260. [Online]. Available: https://www.mdpi.com/2076-3417/13/9/5260.

[24] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “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.

[25] T. Chen, S. Kornblith, M. Norouzi, and G. E. Hinton, “A simple framework for contrastive learning of visual representations,” *CoRR*, vol. abs/2002.05709, 2020. arXiv: 2002.05709. [Online]. Available: https://arxiv.org/abs/2002.05709.

[26] X. Chen, H. Fan, R. B. Girshick, and K. He, “Improved baselines with momentum contrastive learning,” *CoRR*, vol. abs/2003.04297, 2020. arXiv: 2003.04297. [Online]. Available: https://arxiv.org/abs/2003.04297.

[27] X. Chen and K. He, “Exploring simple siamese representation learning,” *CoRR*, vol. abs/2011.10566, 2020. arXiv: 2011.10566. [Online]. Available: https://arxiv.org/abs/2011.10566.

[28] M. Jahromi, P. Buch-Cardona, E. Avots, *et al.*, “Privacy-constrained biometric system for non-cooperative users,” *Entropy*, vol. 21, p. 1033, Oct. 2019. DOI: 10.3390/e21111033.

[29] T. Fawcett, “An introduction to roc analysis,” *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861–874, 2006, ROC Analysis in Pattern Recognition, ISSN: 0167-8655. DOI: `https://doi.org/10.1016/j.patrec.2005.10.010`. [Online]. Available: `https://www.sciencedirect.com/science/article/pii/S016786550500303X`.

[30] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, “Optuna: A next-generation hyperparameter optimization framework,” in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2019.

[31] N. Nachar, “The mann-whitney u: A test for assessing whether two independent samples come from the same distribution,” *Tutorials in Quantitative Methods for Psychology*, vol. 4, Mar. 2008. DOI: `10.20982/tqmp.04.1.p013`.

[32] J. H. Zar, “Spearman rank correlation,” *Encyclopedia of Biostatistics*, vol. 7, 2005.

[33] H. Akoglu, “User’s guide to correlation coefficients,” *Turkish Journal of Emergency Medicine*, vol. 18, Aug. 2018. DOI: `10.1016/j.tjem.2018.08.001`.

[34] O. Majumder, A. Ravichandran, S. Maji, M. Polito, R. Bhotika, and S. Soatto, “Revisiting contrastive learning for few-shot classification,” *CoRR*, vol. abs/2101.11058, 2021. arXiv: `2101.11058`. [Online]. Available: `https://arxiv.org/abs/2101.11058`.

[35] R. R. Selvaraju, A. Das, R. Vedantam, M. Cogswell, D. Parikh, and D. Batra, “Grad-cam: Why did you say that? visual explanations from deep networks via gradient-based localization,” *CoRR*, vol. abs/1610.02391, 2016. arXiv: `1610.02391`. [Online]. Available: `http://arxiv.org/abs/1610.02391`.

A

Class Probability Thresholds

The class probability thresholds used for the five models trained on 5-fold cross-validation are depicted in Table A.1. These thresholds were determined by selecting the point on the ROC curve that has the smallest euclidean distance to the position of true positive rate 1 and false positive rate 0. Plaques are assigned to a class if the predicted probability for that class is the highest and the probability is at least as high as the corresponding threshold. Otherwise, the next highest class is considered. In the case where no class meets their threshold, the class with the highest probability is assigned.

Class	Model 1	Model 2	Model 3	Model 4	Model 5
Diffuse	0.90	0.81	0.93	0.96	0.91
Cored	0.50	0.68	0.58	0.37	0.84
Compact	0.96	0.79	0.67	0.57	0.75
Coarse	0.32	0.87	0.19	0.51	0.60
CAA	0.19	0.16	0.34	0.10	0.15
Subpial	0.76	0.61	0.88	0.64	0.80
OtherAB	0.18	0.05	0.19	0.05	0.11
UndefAB	0.84	0.82	0.98	0.92	0.80
NonAB	0.26	0.18	0.58	0.04	0.06

Table A.1: Class probability thresholds used for each of the five models part of the ensemble.