Predicting Dementia on 3D Brain MRIs

Akhyar Ahmed
8048[illegible]4
ahmed2@u[illegible]am.de

Abstract

Alzheimer's disease (AD) is a neurological condition that causes neurodegeneration and cognitive decline over time. Early detection of the onset of AD is important to initiate treatment and stall progression. Deep learning methods have recently been developed to detect features of AD in structural magnetic images(MRIs) data. Despite the first demonstrations of deep learning applications for brain MRI analysis, people often appraise its outputs as uncertain and difficult to comprehend, hindering the adoption of these algorithms in clinical settings. Another difficulty is detecting patients with mild cognitive impairment (MCI). In this work, I adapt a previously developed deep learning model to classify 3D brain MRIs as AD or Healthy control (HC), with post hoc explanations using layer-wise relevance propagation. I extended this approach to include the intermediate stage of mild cognitive impairment and investigated if the obtained model explanations could find distinct features for this class. Finally, Analyze AD scores between HC, AD, and MCI classes. To evaluate its model quality, overall model performance was assessed using calculated sensitivity, specificity, balanced accuracy, the area under the curve score, and the f1 score. I chose a layer-wise relevance propagator (LRP) and guided backpropagation (GB) in visualization techniques. This analysis yielded several findings: (1) The LRP maps revealed hippocampal and ventricular brain regions, which were established as diagnostic hallmarks to identify dementia on brain MRIs. This result is consistent with the literature. (2) The mean value for the HC class in the AD score distribution of the binary classification model is approximately 0, while the mean value for the AD class is approximately 1. (3) The model's performance decreases in multi-class classification, which implies that the MCI classes are difficult to distinguish from both the HC and AD class. Future work could try to add measured risk factors, such as age, as predictors to see if it can improve the distinction between MCI, AD, and HC.

Keywords: Alzheimer's disease, Dementia, Mild cognitive impairment, Classification.

1 Introduction

Dementia of Alzheimer's type is a chronic neurological disease affecting millions of people worldwide, particularly elderly people. The Yang et al. [2018] also shows around 50 million people globally are affected by dementia, which will be triple by 2050. The disease is characterized by a progressing decline in cognitive function, including memory loss, disorientation, and difficulty in communication, which hinders patients from performing daily activities (Feng et al. [2019];Yang and Mohammed [2020]). Alzheimer's disease is the most common form of dementia, accounting for 60-70% of cases, followed by vascular dementia, Lewy body dementia, and frontotemporal dementia (Yang et al. [2018]). Based on brain cell damage, Alzheimer's disease has mild, moderate, and severe stages.

An early diagnosis of dementia is essential to slow disease progression and give patients adequate care. Unfortunately, diagnosing dementia is challenging because no unique and non-invasive biomarkers

Submitted to Jana Fehr Advanced Machine Learning: Seminar course supervisor (Winter Semester 2022/2023).

allow an easy and low-cost diagnosis. Current guidelines for diagnosing dementia consider a combination of exams, for example, cognitive testing, identifying the protein biomarkers tau and amyloid-beta biomarkers in cerebrospinal fluid (CSF), and brain imaging with MR or PET. Due to this complex diagnostic procedure, reaching a final diagnosis may take between 2.8 to 4.4 years (Gaugler et al. [2022]). To direct efforts toward an early diagnosis of AD, The Alzheimer's Disease Neuroimaging Initiative (ADNI) has been initiated to investigate whether medical imaging, biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early AD. Risk factors that have been associated with the progression of the disease are the genetic factor APOE-ϵ4, and sociodemographic data (Grassi et al. [2019]; Lei et al. [2020]).

Brain imaging is part of the diagnostic procedure for AD, as it gives instant insights into brain morphology and structural changes. Neurodegeneration in the hippocampus and ventricle areas and the primate cortex have been established as imaging biomarkers of dementia (Ávila-Villanueva et al. [2022]). AI is being developed to detect these visual hallmarks automatically. Recent advancements in deep learning and artificial intelligence have developed models to detect features of dementia in brain MRIs automatically (Gassenmaier et al. [2021];Zegers et al. [2021]). The most widely used deep learning method for brain MRI analysis is Convolutional neural networks (CNN). Which use a series of convolutional filters to extract highly predictive feature maps of dementia (Sarraf et al. [2016];Vieira et al. [2017];Litjens et al. [2017]). For example, Sarraf et al. [2016] have implemented an adopted LeNet model and showed that it was, on average 99.999986% accurate in classifying brain MRIs between healthy individuals and individuals with dementia. However, one shortage of using CNNs for this task is that their black-box nature does not explain which features the model has associated with the prediction outcome. Unlike more straightforward learning algorithms, like decision trees, CNNs do not provide an easy-to-understand explanation. Training the multiple parameters (sometimes in the hundreds of thousands) that comprise the complicated architecture of CNNs is required, as their structure includes several layers. In the medical field, making informed decisions for diagnoses and treatments is crucial, rather than relying solely on a binary output from an algorithm. Therefore, if healthcare professionals are to be assisted by CNNs in their daily practice, they must devise methods to interpret and visualize the network's decision.

Several methods have been developed to illustrate the learned predictive features of CNNs. The sensitivity breakdown by Simonyan et al. [2013], guided backpropagation by Springenberg et al. [2014], the deep visualization toolbox of Yosinski et al. [2015] based on regularized optimization, and the deconvolution and occlusion method by Zeiler and Fergus [2013] are some of the most well-known techniques for explaining the predictive outputs of CNNs. These model explanation methods create a unique heatmap per 3D brain MRI. Each highlighted voxel represents the significance of the final classification decision and appears to have the most promise for usage in medical imaging. While heatmaps can provide an intuitive explanation of model predictions, it is important to understand how heatmaps are calculated and what their limits are for any visualization approach that generates them. For instance, strategies based on gradients (e.g., guided backpropagation) in real pictures simply gauge how sensitively the output reacts to changes in the input and may not always correspond with the areas on which the network finds its decision and could potentially also highlight randomly activated areas. Layer-wise relevance propagation (LRP) (Bach et al. [2015]) breaks down the output score of the network (e.g., AD) into the one's pennyworth of the input neurons while preserving the total amount of relevance constant across layers. Compared to gradient and deconvolution methods, LRP has been demonstrated to have better explainability in three natural imaging datasets (Samek et al. [2015]). The LRP method is also used in cognitive neuroscience for single-trial EEG and functional MRI classification, as Thomas et al. [2018] reported. LRP has recently become very popular for explaining clinical disease classification. In an MRI-based Alzheimer's disease classification challenge, Böhle et al. [2019] applied LRP to explain the predictions of a binary model classifying brain MRIs between healthy subjects and AD patients. They showed LRP gives more focused heatmaps than gradient methods. But just like many other works, it focuses only on predicting HC vs. AD but ignores the early spectrum of dementia with mild cognitive impairment.

This work aims to extend the work of Böhle et al. [2019] by extending the predicted outcomes with the class of MCI. Using LRP, I aim to investigate if the prediction model can find different features to predict MCI or AD. To this aim, I adapted the prediction model using 3D brain MRIs from Böhle et al. [2019] and extended it to predict three classes (AD, MCI, HC). I trained the model on 3D images from the Alzheimer's Disease Neuroimaging Initiative (ADNI) (Petersen et al. [2010]) database

and applied LRP to explain the model outcome. I compare model results between the binary and multiclass prediction targets: I demonstrate that the LRP heatmaps successfully illustrated individual contributions to diagnosing AD and may have significant promise as a diagnostic tool. On the other hand, I examined our model's AD score and observed how it changes for both binary and multi-class models.

2 Methodology

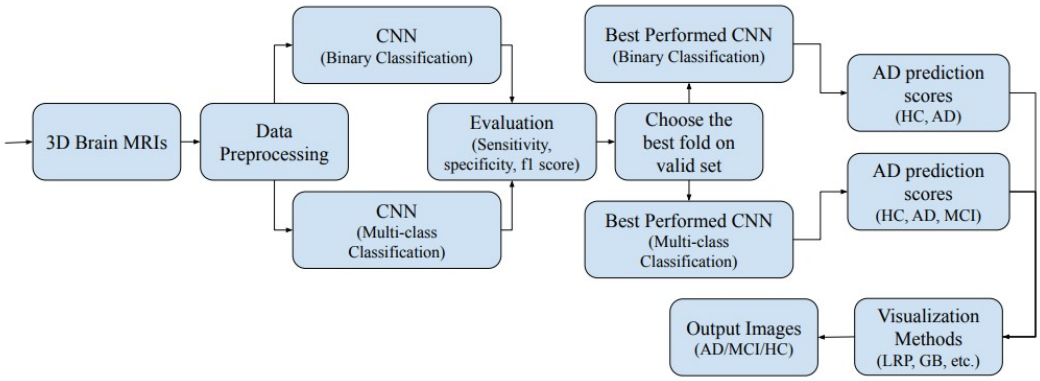

Figure 1: The workflow of this project.

This section describes the experiment setup and workflow(Figure 1). Figure 1 shows the full workflow of our project. I trained models to predict the cognitive status on 3D brain MRIs and implemented two techniques (LRP and GP) to visualize learned predictive features. I evaluated the overall model performance and the feature importance on patient and dataset-level.

2.1 Data preparation and selection

I used the Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset as Böhle et al. [2019] used $n = 939$ 1.5T images of dataset from the same source. For this project, I used the dataset with all the MRIs scanned by 3T scanners which comprised $n = 9219$ MRIs. A number of 2852 MRIs was excluded because some has missing prediction labels, some has file not found error, and some has invalid input error. After the exploratory data analysis, I only took all the samples from AD and HC for binary classification. Later include all the samples from MCI for multi-class classification. Table 1 shows the number of images and patients for the total dataset, the training set and the test set, stratified by cognitive impariment status HC, MCI and AD. Python's NiBabel[1] reads each MRI and transforms them into tensors. Image tensor values were normalized using the Min-max method. Brain regions of interest were segmented and mapped to the brain atlas, using the tool 'Synthseg' (Billot et al. [2023])[2].

Table 1: Dataset and patient distribution.

	Images	HC	MCI	AD	Patients	HC	MCI	AD
All	6367	3099	1720	1548	1204	607	253	344
Train	4109	2011	1094	1004	786	402	160	224
Valid	848	371	277	199	167	80	38	49
Test	1410	716	349	345	251	125	55	71

[1]https://nipy.org/nibabel/

[2]https://surfer.nmr.mgh.harvard.edu/fswiki/SynthSeg

2.2 Model architecture

CNNs are a type of neural network specifically designed for processing array data such as pictures and videos (LeCun et al. [2015]). They have several hidden layers, including the input and output layers. I employed the same CNN architecture as Böhle et al. [2019]. In addition, I changed the output feature into three for the multi-class classification model, as it gets data for HC, AD, and MCI class labels. For binary classification, the output feature remains the same. The same optimizer is used, which is Adam (Kingma and Ba [2017]) optimizer. All models were trained four times. The repetition with the highest balanced accuracy was applied to predict cognitive outcomes in the test set. Later, I used a density plot for each model to compare the distribution of AD scores among each class. The Gaussian kernel (Chung [2020]) method is used to smoothen these plots.

2.3 Visualization techniques

2.3.1 Layer-wise relevance propagation

The core method to explain model predictions is LRP (Bach et al. [2015]). The LRP method assigns significance to individual input nodes by iteratively tracing contributions to the final output node. Several types of LRP algorithms are available that share the same idea of preserving the total relevance of class label, such as the activation strength of an output node for a certain class per layer. Böhle et al. [2019] used the β-rule of LRP:

$$R_{l,l+1}^{i \to j} = ((1+\beta)\frac{z_{ij}^{+}}{z_j^{+}} - \beta\frac{z_{ij}^{-}}{z_j^{-}})R_{l+1}^{j}$$

Here, $z_{ij}^{+/-}$ represents the amount of positive/negative input that node i contributed to node j. The respective contributions are divided by the sum over all positive/negative contributions of the nodes in the layer l, $z_j^{+/-} = \sum_i z_{ij}^{+/-}$ such that the relevance is maintained from layer $l+1$ to layer l. β value controls which contribution should visualize in the heatmap (Böhle et al. [2019]). The zero β value will show positive contributions in the heatmap, but non-zero β values also account for the inhibitory effects of neuron activations. I applied the β-values $0, 0.5$, and 1 to investigated how they influenced the resulting heatmaps. In this project, I adapted the implementation of the LRP method for predicting the cognitive status by Böhle et al. [2019].

2.3.2 Guided backpropagation

I also adapted the implementation of the GBP visualization method from Böhle et al. [2019] to compare heatmaps from the LRP to a gradient-based visualization method. The GBP is a gradient-based visualization approach that visualizes the gradient concerning pictures when backpropagating through the Relu activation function (Springenberg et al. [2014]). Only positive gradients can flow in this method while it sets the negative gradient to zero. Forward pass is,

$$f_i^{l+1} = ReLu(f_i^l, 0)$$

Here, f is the feature map, l is a layer. And the backward pass is:

$$R_i^l = (f_i^l > 0).(R_i^{l+1} > 0).(R_i^{l+1})$$

Where R holds the backpropagation result of layer l.

2.4 Evaluation metrics

The discriminatory performance of prediction models was assessed using the metrics Sensitvity, Specificity, Balance accuracy, F1 score, and Area Under the Curve (AUC) score in receiver operating characteristic (ROC). Sensitivity measures the correctness of detecting the positive class of a model. Specificity shows how correctly a model detects negative classes. The average of sensitivity and specificity is called balanced accuracy, by which we can tell the average accuracy over minority and majority classes. The F1 score is a combination of precision and recall. The F1 score tells us the frequency of correctly predicted classes over the dataset. The AUC score shows how well a model correctly distinguishes a positive class from a negative one.

Table 2: Classification model results on evaluation metrics

	Fold	Sensitivity	Specificity	Balanced accuracy	F1 score
Binary classification	0	0.78	0.81	**0.89**	0.88
	1	0.30	**0.92**	0.87	0.89
	2	**0.80**	0.74	0.83	0.88
	3	0.48	0.90	0.87	**0.90**
Multi-class classification	0	**0.37**	0.78	**0.62**	**0.61**
	1	0.18	**0.89**	0.59	**0.61**
	2	0.26	0.83	0.56	0.57
	3	0.27	0.83	0.60	0.60

2.5 Atlas-based importance metrics

To answer the question, "Can this model activate relevant scores of different brain areas?" I used the same atlas-based importance metrics which Böhle et al. [2019] used in their experiment.

Sum of AD importance per area Simply sums all the generated LRP values for each region of each pixel value. This metric generally represents which region of the brain was relevant for the model to predict AD.

Size-normalized AD importance metric This metric represents the regional mean relevance or susceptibility of AD importance area. It can be calculated from a division between the sum of the relevance score per zone and the size of the zone. Low values across wide regions may simply result from statistical fluctuations in the data. At the same time, clusters of relevance (LRP) or susceptibility (GB) in a small area may indicate systematic associations of brain regions with cognitive status.

Gain – ratio of values concerning the average HC The 'gain' of relevance was assessed for each pixel value to verify that the LRP-algorithm did not interpret AD-relevant regions as 'relevance-free" in negative health controls (HC). The gain is represented by the ratio to the average HC in that particular area. The "gain" will emphasize the places where the two scenarios most significantly diverge.

3 Results

Section 3.1 shows our evaluation metrics' results and compares the scores of each fold (see Table 2). Later, AD score distribution showed among all the classes for binary and multi-class classification methods. Section 3.2 plots the LRP heatmaps of two individual patients and compares the results for different beta values. 3.3 shows the heatmaps created on the test set from the different beta values. Finally, section 3.4 compares the heatmaps quantitatively using several atlas-based importance metrics.

3.1 Performance to predict cognitive status on brain MRIs

Table 2, binary classification part shows each fold performance result for the binary model. The multi-class classification part shows the same for the multi-class model. Table 2 shows that fold-2 and fold-0 checkpoints can correctly predict the most positive classes for the binary and multi-class models. Where fold-1 of both models shows the best result in correctly predicting negative classes. If we consider the balance of sensitivity and specificity, then fold-0 of both models shows the best balance accuracy score, 89%, and 62%, respectively. Fold-3 is the binary model's most frequent checkpoint in predicting positive class to the negative, whereas fold-0 and fold-1 checkpoints are equally frequent for the multi-class model.

Figure 2 shows the binary (left side) and multi-class (right one) classification models' AUC scores in the ROC curve. Both ROC curves show how well both models distinguish the positive class from negative between HC vs. AD, HC vs. AD+MCI, and HC vs. MCI.

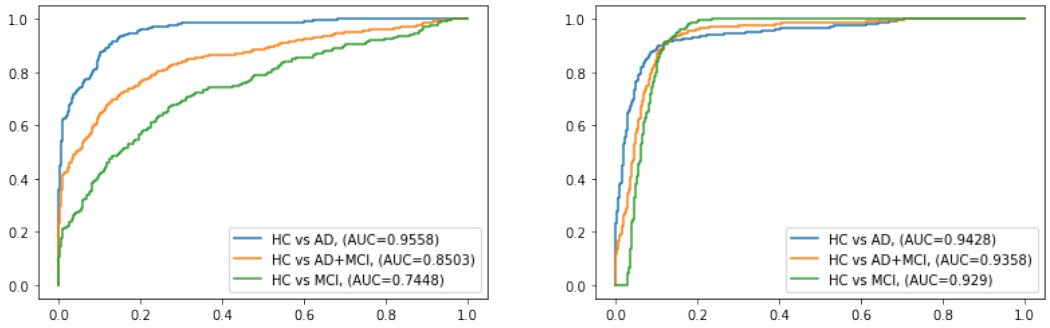

Figure 2: ROC curve and AUC score comparison between binary classification(left) and multi-class classification(right).

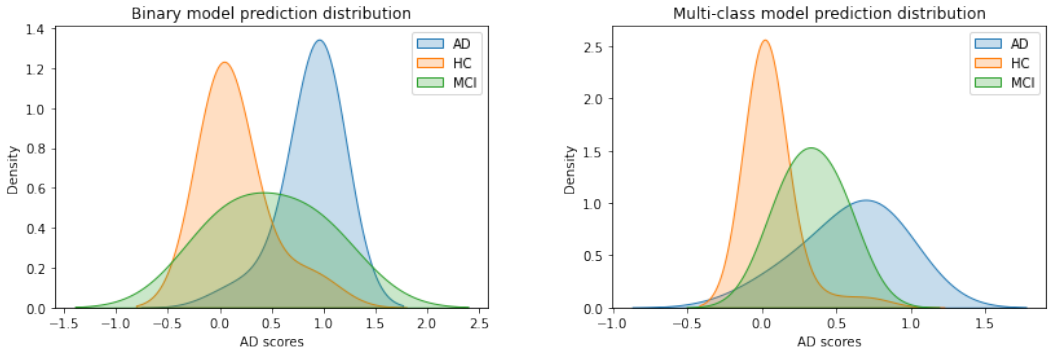

Figure 3: AD score distributions.

Figure 3 has two density plots. The binary model's AD-score distribution is shown on the left, whereas the multi-class model's prediction distribution is shown on the right. For the binary model, the mean point for HC, MCI, and AD classes are 0, 0.5, and 1, respectively. And for the multi-class model, the mean point for HC, MCI, and AD classes are 0, 0.4, and 0.8, respectively.

3.2 Neurological relevance

This part shows the result of individual heatmaps. Böhle et al. [2019] showed three slices of heatmaps for randomly picked two AD patients to check the relevance and diversity of the heatmap (see Figure 24). I extended it and plotted the binary and multi-class models for all β-values (see all the Figures 24, 25, 26, 27, 28, 29).I stacked the middle slice image pairs for different beta values and merged them into two figures (see Figure 4).We can easily identify the relevant regions for detecting the class of binary(left) and multi-class(right) models when the β-value is zero. But when we increased the β-value to 0.5, we can see that (Figure 4), the previously activated region becomes more relevant for both models and gets scattered from its original area. β-value = 1 also showed the same behavior in both models.

For each patient, the sum of relevance for each region is also shown in Figure 21. The Figure 21 clearly shows that the locations that most influenced the network decision for the two patients are rather different. The first patient (Patient A) has several influenced regions, e.g., Right Cerebral White Matter, Right Lateral Ventricle, Right Cerebellum Cortex, etc. In the second patient (Patient B), Right Ventral DC, Right Putamen, and Right Thalamus are some of the most relevant regions for the AD class.

3.3 Explaining predictive features using heatmaps calculated with LRP and GB

Böhle et al. [2019] showed the average heatmaps for AD patients and HCs, separately for LRP with different β-values (β = 0, 0.5, 1) and GB (Figure 6, 7, 8, 5). Here I extended it to a multi-class

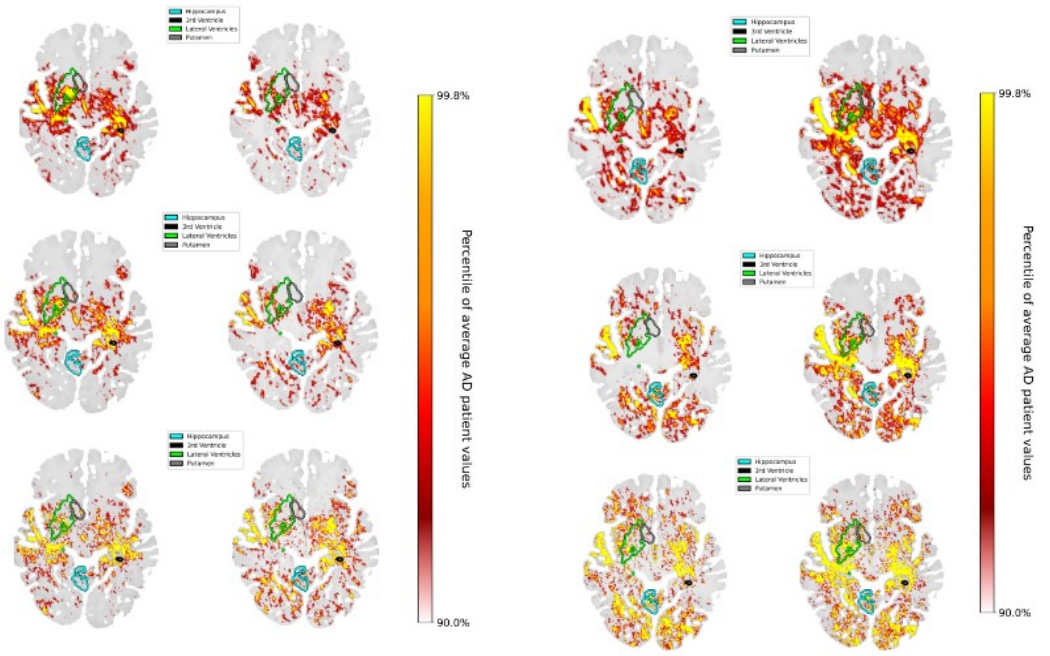

Figure 4: LRP activations in central brain slices of two patients from binary classification (left) and multi-class classification(right) for different $\beta = 0(top), 0.5(middle), 1(bottom)$ values.

classification model (see Figure 12, 13, 14, 11). If we compare the LRP and GB heatmaps, there was not much change in the activated relevant area. Though LRP heatmaps are more specific and easy to explainable. Whereas GB's heatmap looks suspicious as it can only activate the left part of the brain images. I combined all the different β-values heatmaps for both models (see Figure 9,15) and saw that for every β-value all the slices are relatively same though Binder et al. [2016] stated that higher the β-value larger the inhibitory contribution. Böhle et al. [2019] and this experiment focused on positive contribution of AD that is why all the evaluation held with β-value 0.

3.4 Atlas-based importance metrics

Figures 19 and 20 show the AD importance sum per area for LRP and GB. In both models, LRP is implemented with β=0. The size of the respective brain area also dominates this metric's feature (Böhle et al. [2019]). If we see both models' LRP results in Figures 19 and 20, there is consistently a huge gap between HC and AD in calculating mean importance values per area. But the significant part is only the right side of the brain shows perfect results. In contrast, the left part of the brain is completely deactivated.

In the size-normalized AD importance metric, The normalization of the total sum of importance is done based on the size of the brain area, which highlights the difference in distributions between HCs and AD patients. This difference is more visible in LRP than GB, as the distributions are highly overlapping for GB but not for LRP. While several subtypes of AD have been studied recently, which led us to that the network has mastered the art of differentiating between them and based its decisions on various structural factors for various individuals (Ferreira et al. [2017];Park et al. [2017]).

Figure 22, 23, shows the gain metric outcomes for both true positive and true negative scenarios. This measure again underlines the greater separation between AD patients and HCs under the LRP method and enables the presentation of the LRP and GB findings on the same scale. Most gain for LRP in binary model has been found in areas of Left Ventral DC, Right Amygdala, Left Accumbens Area (Figure 22) and for Multi-class model Left Amygdala, Left Hippocampus has the most gain ratio for LRP (Figure 23).

4 Discussion

In this project, I inherited Böhle et al. [2019] method and extended it to another class of mild empirical impairment (MCI). I wanted to see how the prediction score behaves in the AD detection task and also tried to explain the AD class using LRP. In the performance part, the binary model hit the same level as Böhle et al. [2019], which was obvious, but when I sent some unknown features (MCI) to predict, its performance started declining. And it became more difficult when I introduced MCI class as a model class in the model. It also declined the prediction scores for other classes (Figure 3). One takeaway is that MCI is a fragile feature to predict AD. Because it consistently shows AD less than or equal to 0.5. Later plotting two patients' heatmaps revealed that the LRP method could highlight AD's important hallmarks (Figure 4). I examined the heatmaps with various values and categorization categories at the population level, including AD patients, HCs, MCIs, true positives, and false positives. Three alternative significance metrics were used to assess the relevance of the brain areas, including the total importance per area, the size-normalized AD importance, and the gain as a ratio between AD and HC. Moreover, the relevant regions identified by the LRP heatmaps align with the findings reported in the literature. Notably, expert annotations on biomarkers were not required during the training process to generate these LRP heatmaps. Using LRP for comprehensive network analysis and a straightforward classification task could be a potential diagnostic tool for healthcare professionals and enhance trust in computer-aided diagnoses by offering an understandable explanation for the decision.

While LRP heatmaps show potential for visualizing neural network decisions, it is important to acknowledge certain limitations of this method and other heatmap techniques in the context of this study. Firstly denoting the area of interest in our brain MRIs. Though people use the same dataset, sometimes people use different dimension data for which mask is unavailable. For me generating a neurodegenerative atlas is very difficult to use. And for such a task, we need masks to train our CNNs. Secondly, biased on ADNI, more diverse datasets with larger sample sizes may help to improve models.Third, The LRP method is highly dependable in $beta$-value, which is not tunned in our case. This study didn't face such an obstacle, but it must be handled. Finally, lack of pre-knowledge about brain cells. I would say this is my only personal limitation in this project.

β is a core parameter for calculating LRP. But there is no specific method to hyper-tune it. In classification tasks, people examine the resilience for different values of β. And in this analysis, changing the β-value changed the result relatively. So I believe a proper tuned β-value surely increase the result of this project. Second improvement sector is the feature MCI. If we could know more feature about MCI than I am sure it will improve the overall prediction of both multi-class and binary classification model.

References

Marina Ávila-Villanueva, Alberto Marcos Dolado, Jaime Gómez-Ramírez, and Miguel Fernández-Blázquez. Brain structural and functional changes in cognitive impairment due to alzheimer's disease. *Front. Psychol.*, 13:886619, June 2022.

Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLoS ONE*, 10, 2015.

Benjamin Billot, Douglas N. Greve, Oula Puonti, Axel Thielscher, Koen Van Leemput, Bruce Fischl, Adrian V. Dalca, and Juan Eugenio Iglesias. SynthSeg: Segmentation of brain MRI scans of any contrast and resolution without retraining. *Medical Image Analysis*, 86:102789, may 2023. doi: 10.1016/j.media.2023.102789. URL `https://doi.org/10.1016%2Fj.media.2023.102789`.

Alexander Binder, Sebastian Bach, Grégoire Montavon, Klaus-Robert Müller, and Wojciech Samek. Layer-wise relevance propagation for deep neural network architectures. In Kuinam J. Kim and Nikolai Joukov, editors, *Information Science and Applications (ICISA) 2016*, volume 376 of *Lecture Notes in Electrical Engineering*, pages 913–922. Springer Singapore, Singapore, 2016. ISBN 978-981-10-0557-2. doi: 10.1007/978-981-10-0557-2_87. URL `http://dx.doi.org/10.1007/978-981-10-0557-2_87`.

Moritz Böhle, Fabian Eitel, Martin Weygandt, and Kerstin Ritter. Layer-wise relevance propagation for explaining deep neural network decisions in mri-based alzheimer's disease classification.

Frontiers in Aging Neuroscience, 11, 2019. ISSN 1663-4365. doi: 10.3389/fnagi.2019.00194. URL https://www.frontiersin.org/articles/10.3389/fnagi.2019.00194.

Moo K. Chung. Gaussian kernel smoothing. *ArXiv*, abs/2007.09539, 2020.

Chiyu Feng, A. Elazab, Peng Yang, Tianfu Wang, Feng Zhou, Huoyou Hu, Xiaohua Xiao, and Baiying Lei. Deep learning framework for alzheimer's disease diagnosis via 3d-cnn and fsbi-lstm. *IEEE Access*, 7:63605–63618, 2019.

Daniel Ferreira, Chloë Verhagen, Juan Andrés Hernández-Cabrera, Lena Cavallin, Chun-Jie Guo, Urban Ekman, J-Sebastian Muehlboeck, Andrew Simmons, José Barroso, Lars-Olof Wahlund, and Eric Westman. Distinct subtypes of alzheimer's disease based on patterns of brain atrophy: longitudinal trajectories and clinical applications. *Sci. Rep.*, 7:46263, April 2017.

Sebastian Gassenmaier, Thomas Küstner, Dominik Nickel, Judith Herrmann, Rüdiger Hoffmann, Haidara Almansour, Saif Afat, Konstantin Nikolaou, and Ahmed E Othman. Deep learning applications in magnetic resonance imaging: Has the future become present? *Diagnostics (Basel)*, 11(12):2181, November 2021.

Joseph Gaugler, Bryan James, Tricia Johnson, Jessica Reimer, Michele Solis, Jennifer Weuve, Rachel F. Buckley, and Timothy J. Hohman. 2022 alzheimer's disease facts and figures. *Alzheimer's & Dementia*, 18(4):700–789, 2022. doi: https://doi.org/10.1002/alz.12638. URL https://alz-journals.onlinelibrary.wiley.com/doi/abs/10.1002/alz.12638.

Massimiliano Grassi, Nadine Rouleaux, Daniela Caldirola, David Loewenstein, Koen Schruers, Giampaolo Perna, Michel Dumontier, and Alzheimer's Disease Neuroimaging Initiative . A novel ensemble-based machine learning algorithm to predict the conversion from mild cognitive impairment to alzheimer's disease using socio-demographic characteristics, clinical information, and neuropsychological measures. *Frontiers in Neurology*, 10, 2019. ISSN 1664-2295. doi: 10.3389/fneur.2019.00756. URL https://www.frontiersin.org/articles/10.3389/fneur.2019.00756.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2017.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, May 2015.

Baiying Lei, Mengya Yang, Peng Yang, Feng Zhou, Wen Hou, Wenbin Zou, Xia Li, Tianfu Wang, Xiaohua Xiao, and Shuqiang Wang. Deep and joint learning of longitudinal data for alzheimer's disease prediction. *Pattern Recognition*, 102:107247, 2020. ISSN 0031-3203. doi: https://doi.org/10.1016/j.patcog.2020.107247. URL https://www.sciencedirect.com/science/article/pii/S0031320320300534.

Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen A.W.M. van der Laak, Bram van Ginneken, and Clara I. Sánchez. A survey on deep learning in medical image analysis. *Medical Image Analysis*, 42:60–88, 2017. ISSN 1361-8415. doi: https://doi.org/10.1016/j.media.2017.07.005. URL https://www.sciencedirect.com/science/article/pii/S1361841517301135.

Jong-Yun Park, Han Kyu Na, Sungsoo Kim, Hyunwook Kim, Hee Jin Kim, Sang Won Seo, Duk L Na, Cheol E Han, Joon-Kyung Seong, and Alzheimer's Disease Neuroimaging Initiative. Robust identification of alzheimer's disease subtypes based on cortical atrophy patterns. *Sci. Rep.*, 7:43270, March 2017.

R C Petersen, P S Aisen, L A Beckett, M C Donohue, A C Gamst, D J Harvey, C R Jack, Jr, W J Jagust, L M Shaw, A W Toga, J Q Trojanowski, and M W Weiner. Alzheimer's disease neuroimaging initiative (ADNI): clinical characterization. *Neurology*, 74(3):201–209, January 2010.

Wojciech Samek, Alexander Binder, Grégoire Montavon, Sebastian Bach, and Klaus-Robert Müller. Evaluating the visualization of what a deep neural network has learned. *CoRR*, abs/1509.06321, 2015. URL http://arxiv.org/abs/1509.06321.

Saman Sarraf, Ghassem Tofighi, and . Deepad: Alzheimer's disease classification via deep convolutional neural networks using mri and fmri. *bioRxiv*, 2016. doi: 10.1101/070441. URL https://www.biorxiv.org/content/early/2016/08/21/070441.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *CoRR*, abs/1312.6034, 2013.

Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin A. Riedmiller. Striving for simplicity: The all convolutional net. *CoRR*, abs/1412.6806, 2014.

Armin W. Thomas, Hauke R. Heekeren, Klaus-Robert Müller, and Wojciech Samek. Interpretable lstms for whole-brain neuroimaging analyses. *CoRR*, abs/1810.09945, 2018. URL http://arxiv.org/abs/1810.09945.

Sandra Vieira, Walter H.L. Pinaya, and Andrea Mechelli. Using deep learning to investigate the neuroimaging correlates of psychiatric and neurological disorders: Methods and applications. *Neuroscience Biobehavioral Reviews*, 74:58–75, 2017. ISSN 0149-7634. doi: https://doi.org/10.1016/j.neubiorev.2017.01.002. URL https://www.sciencedirect.com/science/article/pii/S0149763416305176.

Chengliang Yang, Anand Rangarajan, and Sanjay Ranka. Visual explanations from deep 3d convolutional neural networks for alzheimer's disease classification. *CoRR*, abs/1803.02544, 2018. URL http://arxiv.org/abs/1803.02544.

Kuo Yang and Emad A. Mohammed. A review of artificial intelligence technologies for early prediction of alzheimer's disease. *ArXiv*, abs/2101.01781, 2020.

Jason Yosinski, Jeff Clune, Anh M Nguyen, Thomas J. Fuchs, and Hod Lipson. Understanding neural networks through deep visualization. *ArXiv*, abs/1506.06579, 2015.

C.M.L. Zegers, J. Posch, A. Traverso, D. Eekers, A.A. Postma, W. Backes, A. Dekker, and W. van Elmpt. Current applications of deep-learning in neuro-oncological mri. *Physica Medica*, 83: 161–173, 2021. ISSN 1120-1797. doi: https://doi.org/10.1016/j.ejmp.2021.03.003. URL https://www.sciencedirect.com/science/article/pii/S1120179721001198.

Matthew D. Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European Conference on Computer Vision*, 2013.

A Appendix

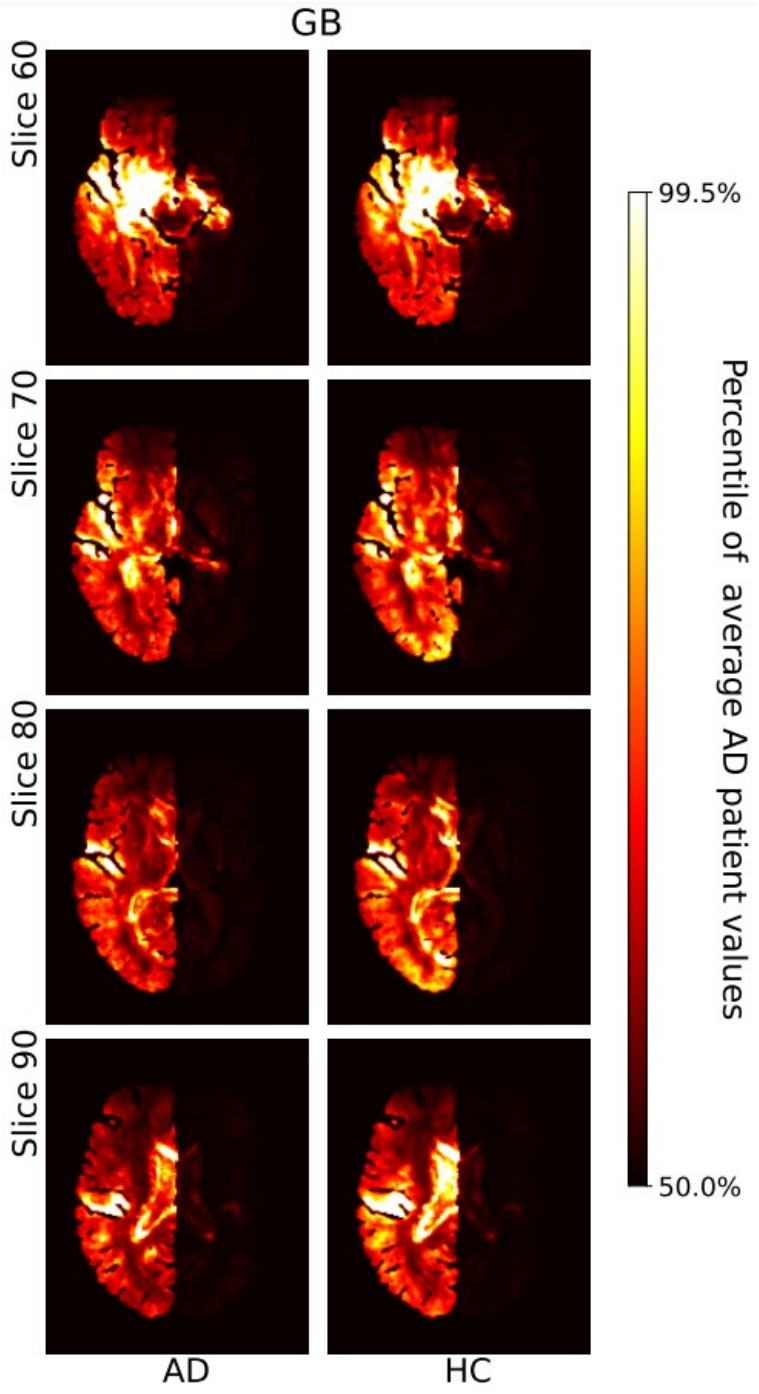

Figure 5: Average heatmaps for AD patients and healthy controls (HCs) in the test set are shown for GB (right) in binary model. The values in the average heatmaps that are higher than the 50th percentile and lower than the 99.5th percentile are linearly color-coded as shown on the scale. Values below (above) these numbers are black (white).

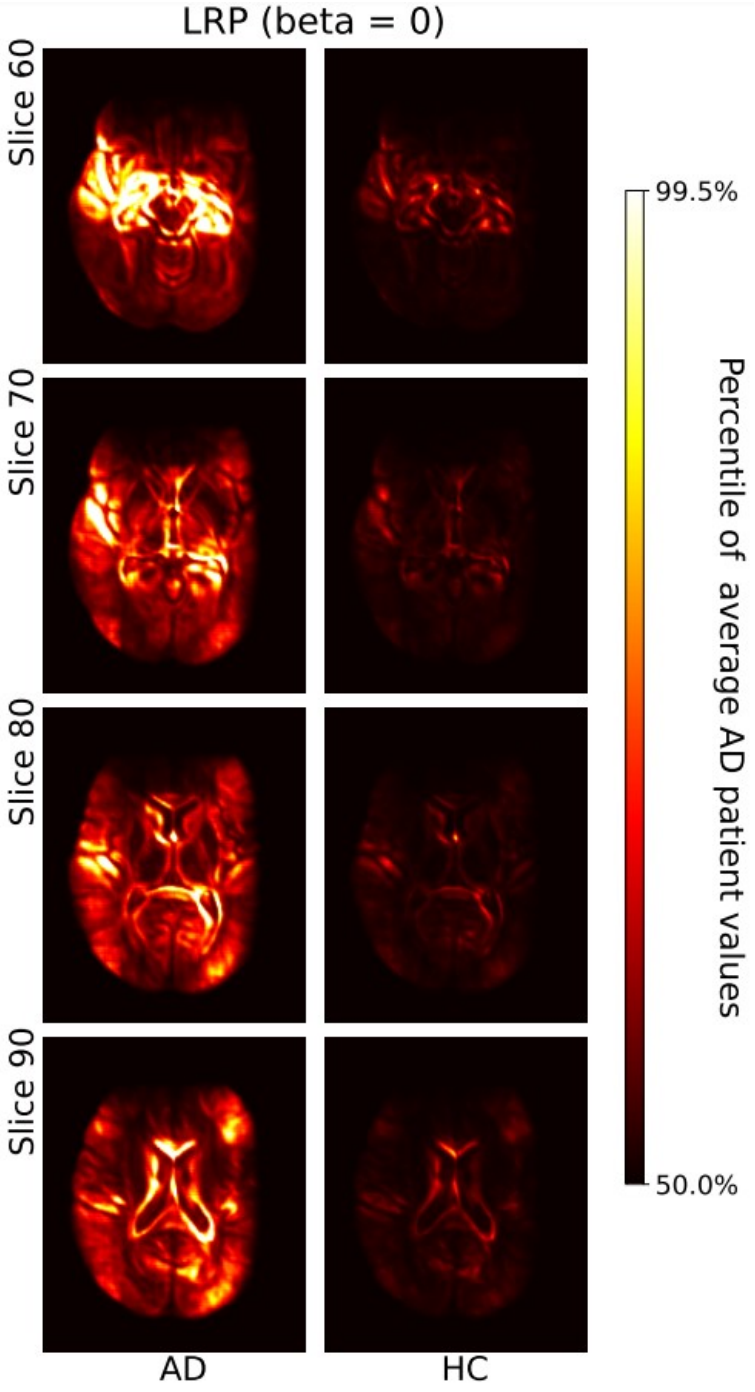

Figure 6: Average heatmaps for AD patients and healthy controls (HCs) in the test set are shown for LRP with = 0 in binary model. The values in the average heatmaps that are higher than the 50th percentile and lower than the 99.5th percentile are linearly color-coded as shown on the scale. Values below (above) these numbers are black (white).

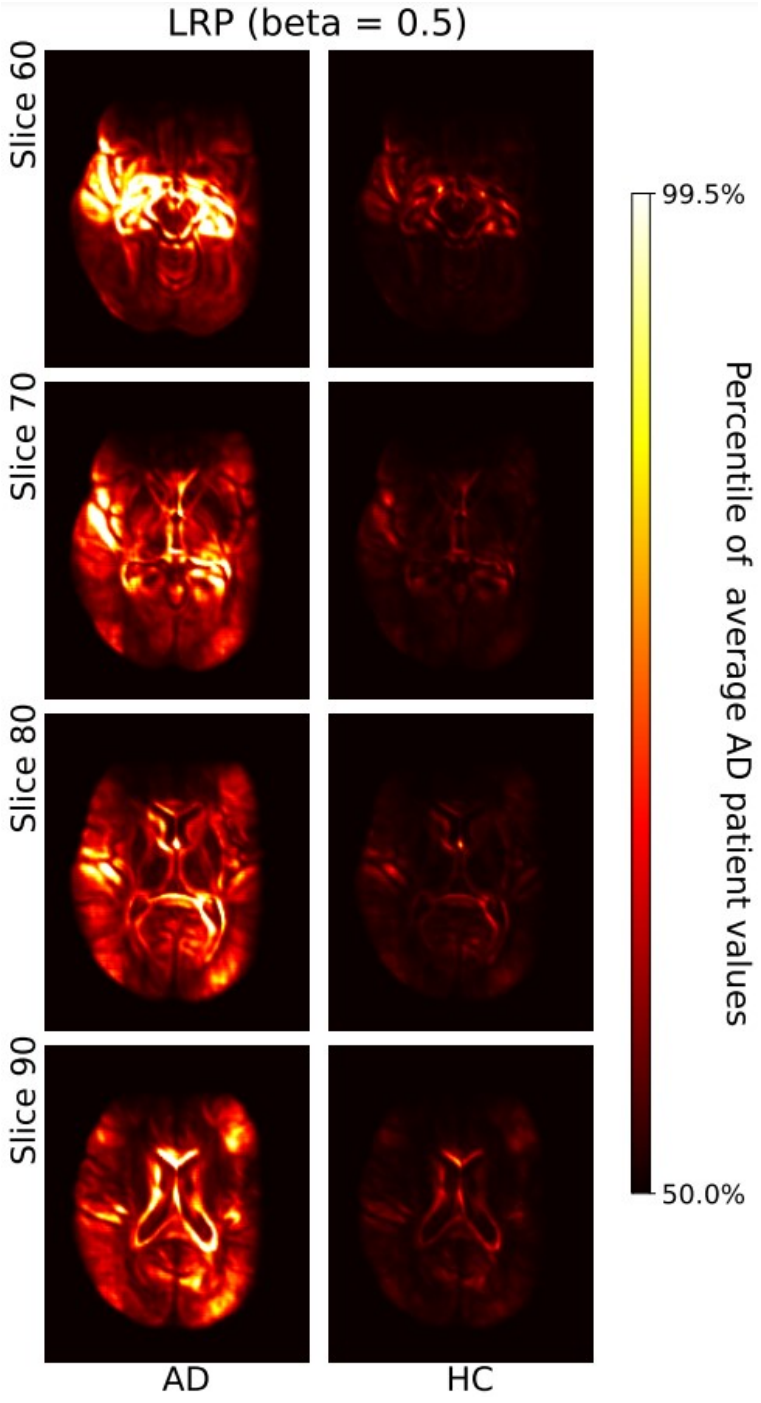

Figure 7: Average heatmaps for AD patients and healthy controls (HCs) in the test set are shown for LRP with = 0.5 in binary model. The values in the average heatmaps that are higher than the 50th percentile and lower than the 99.5th percentile are linearly color-coded as shown on the scale. Values below (above) these numbers are black (white).

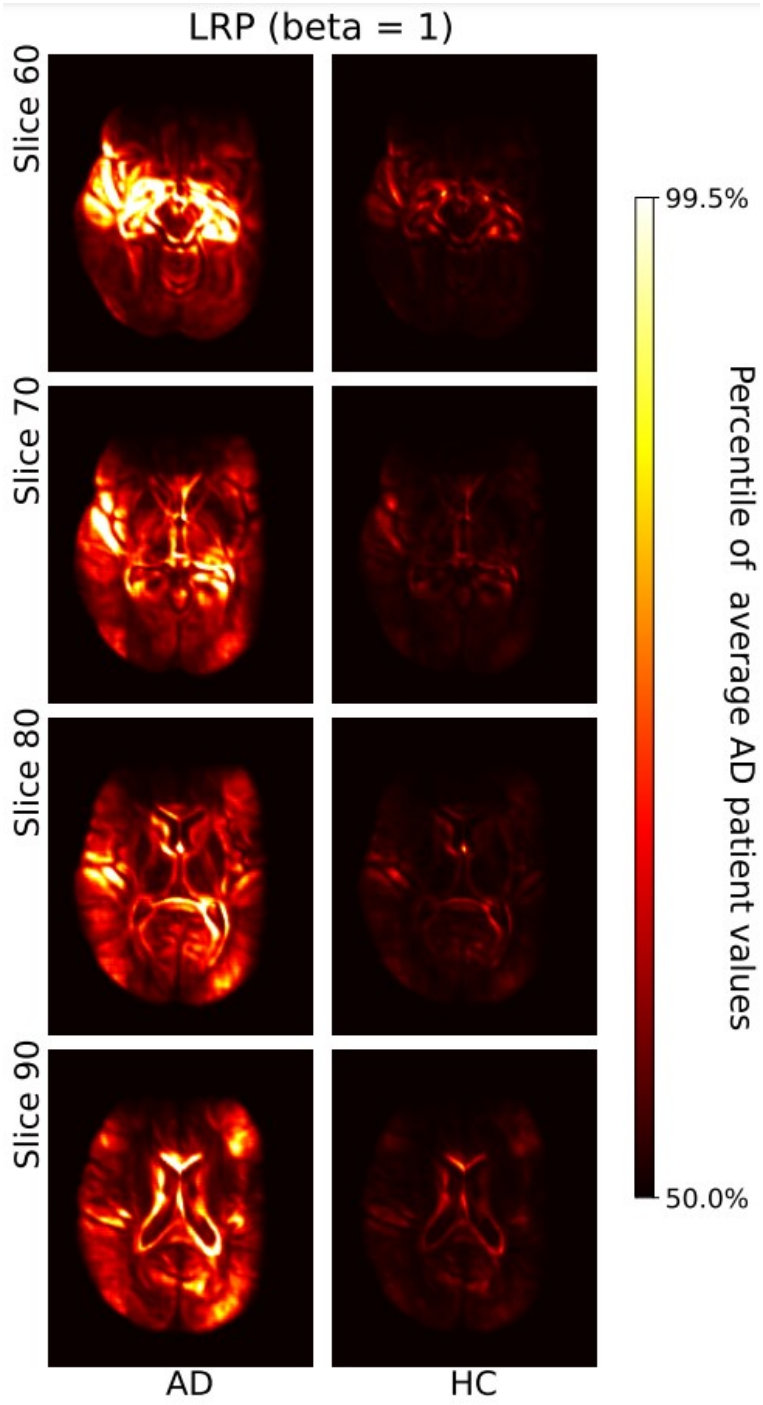

Figure 8: Average heatmaps for AD patients and healthy controls (HCs) in the test set are shown for LRP with = 1 in binary model. The values in the average heatmaps that are higher than the 50th percentile and lower than the 99.5th percentile are linearly color-coded as shown on the scale. Values below (above) these numbers are black (white).

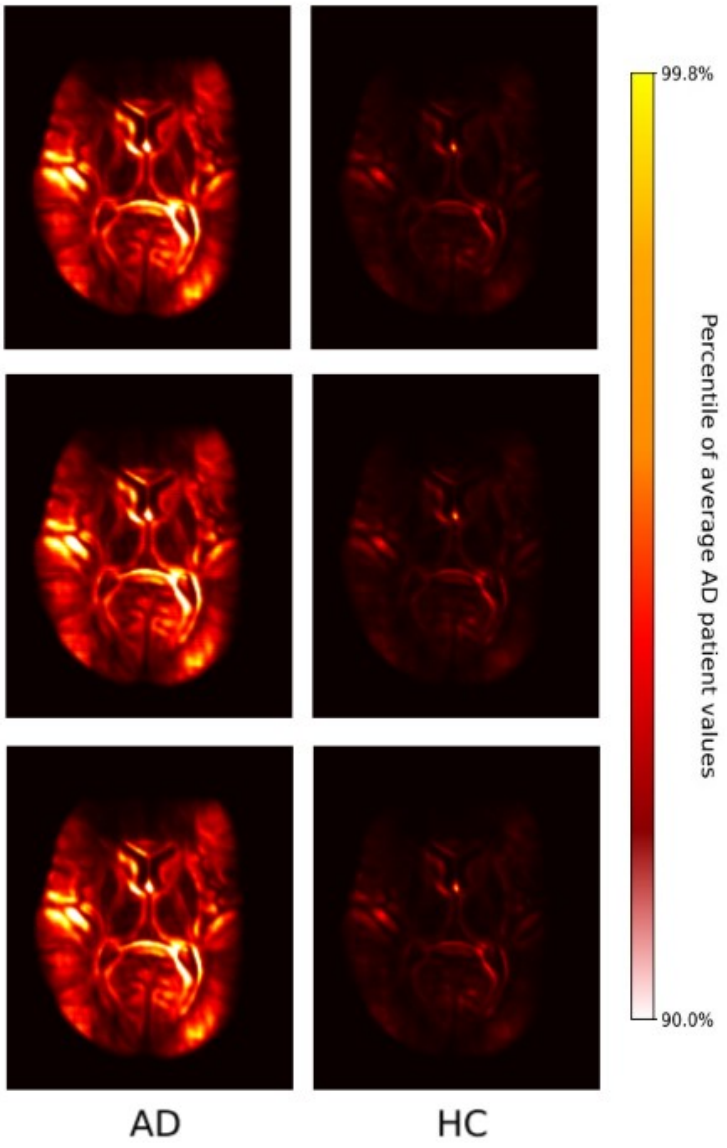

Figure 9: Average heatmaps for AD patients and healthy controls (HCs) for slice 80 are shown seperately for LRP with = 0 (top slice), 0.5(Middle slice), 1 (bottom slice) in binary model. The values in the average heatmaps that are higher than the 90th percentile and lower than the 99.8th percentile are linearly color-coded as shown on the scale. Values below (above) these numbers are black (white).

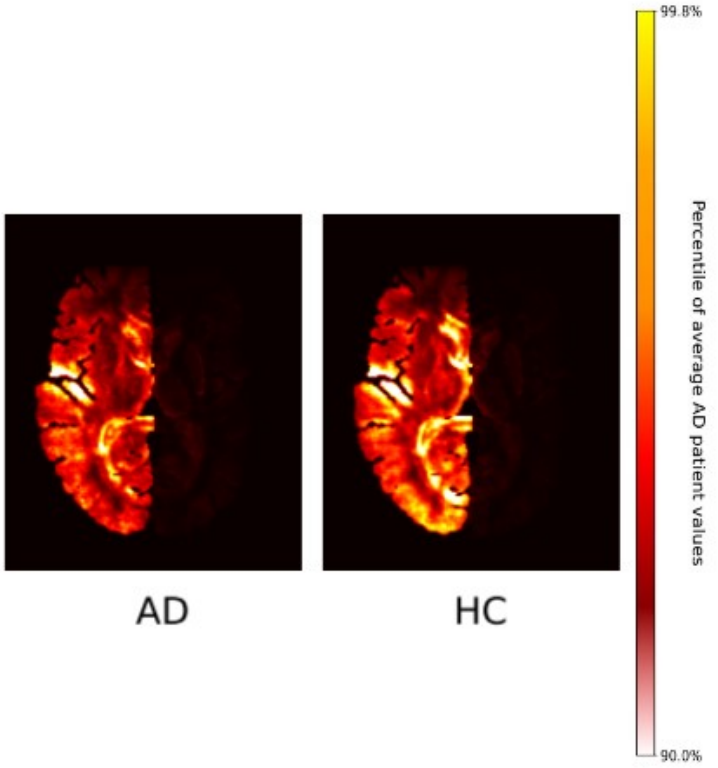

Figure 10: Average heatmaps for AD patients and healthy controls (HCs) in a single slice (80) is shown for GB in binary model. The values in the average heatmaps that are higher than the 90th percentile and lower than the 99.8th percentile are linearly color-coded as shown on the scale. Values below (above) these numbers are black (white).

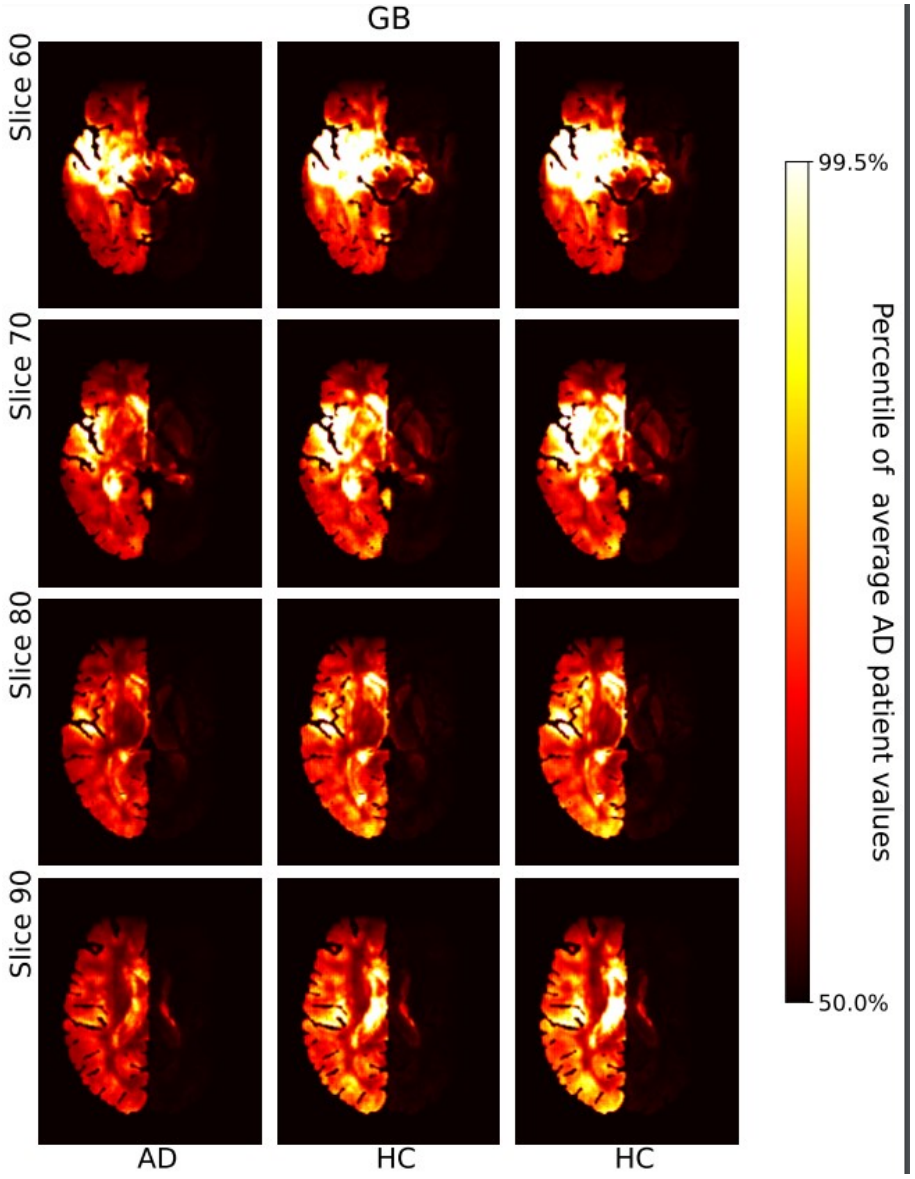

Figure 11: Average heatmaps for AD patients, MCI, and healthy controls (HCs) in the test set are shown for GB in multi-class model. The values in the average heatmaps that are higher than the 50th percentile and lower than the 99.5th percentile are linearly color-coded as shown on the scale. Values below (above) these numbers are black (white).

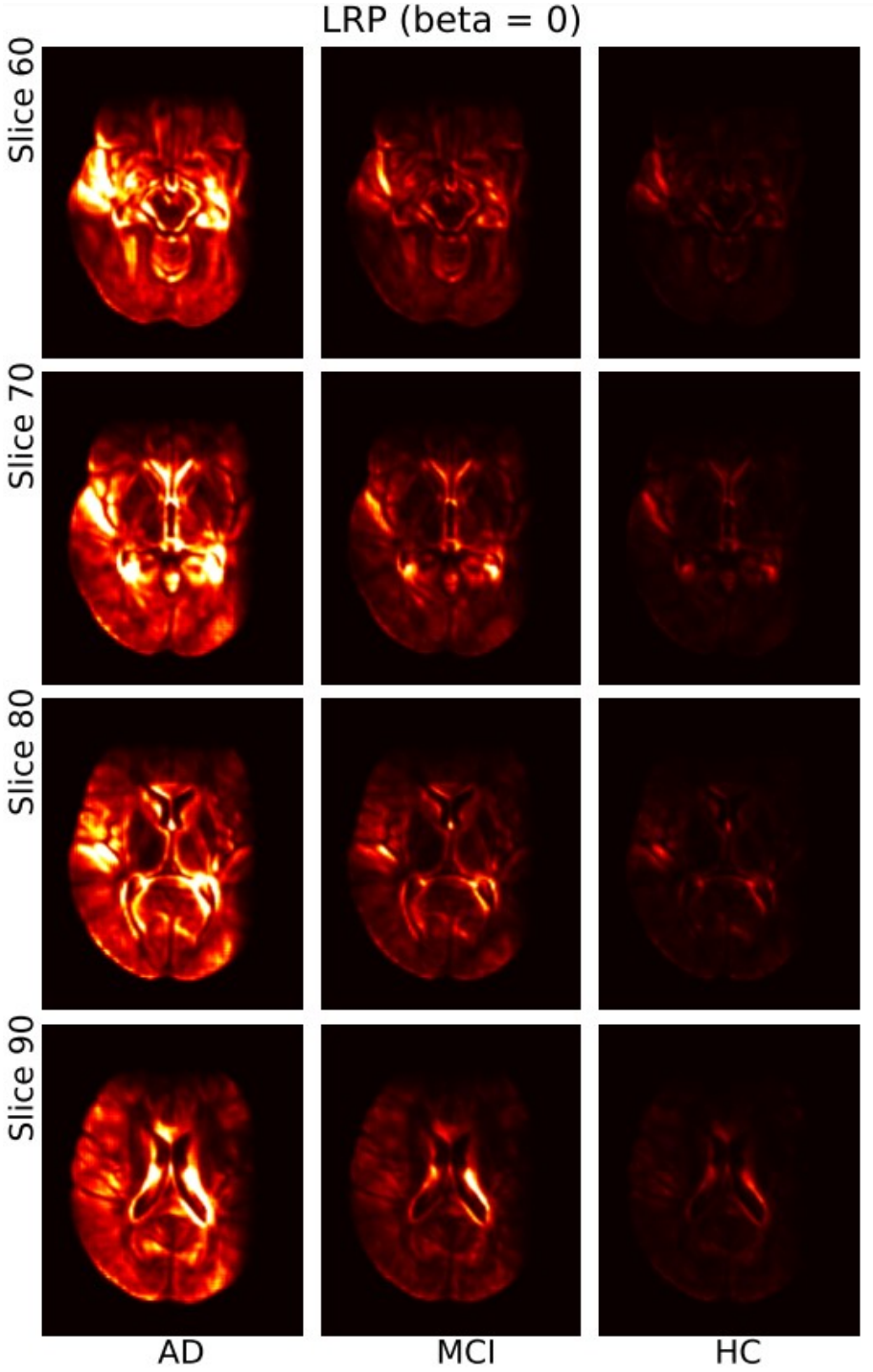

Figure 12: Average heatmaps for AD patients, MCI, and healthy controls (HCs) in the test set are shown for LRP with = 0 in multi-class model.The values in the average heatmaps that are higher than the 50th percentile and lower than the 99.5th percentile are linearly color-coded as shown on the scale. Values below (above) these numbers are black (white).

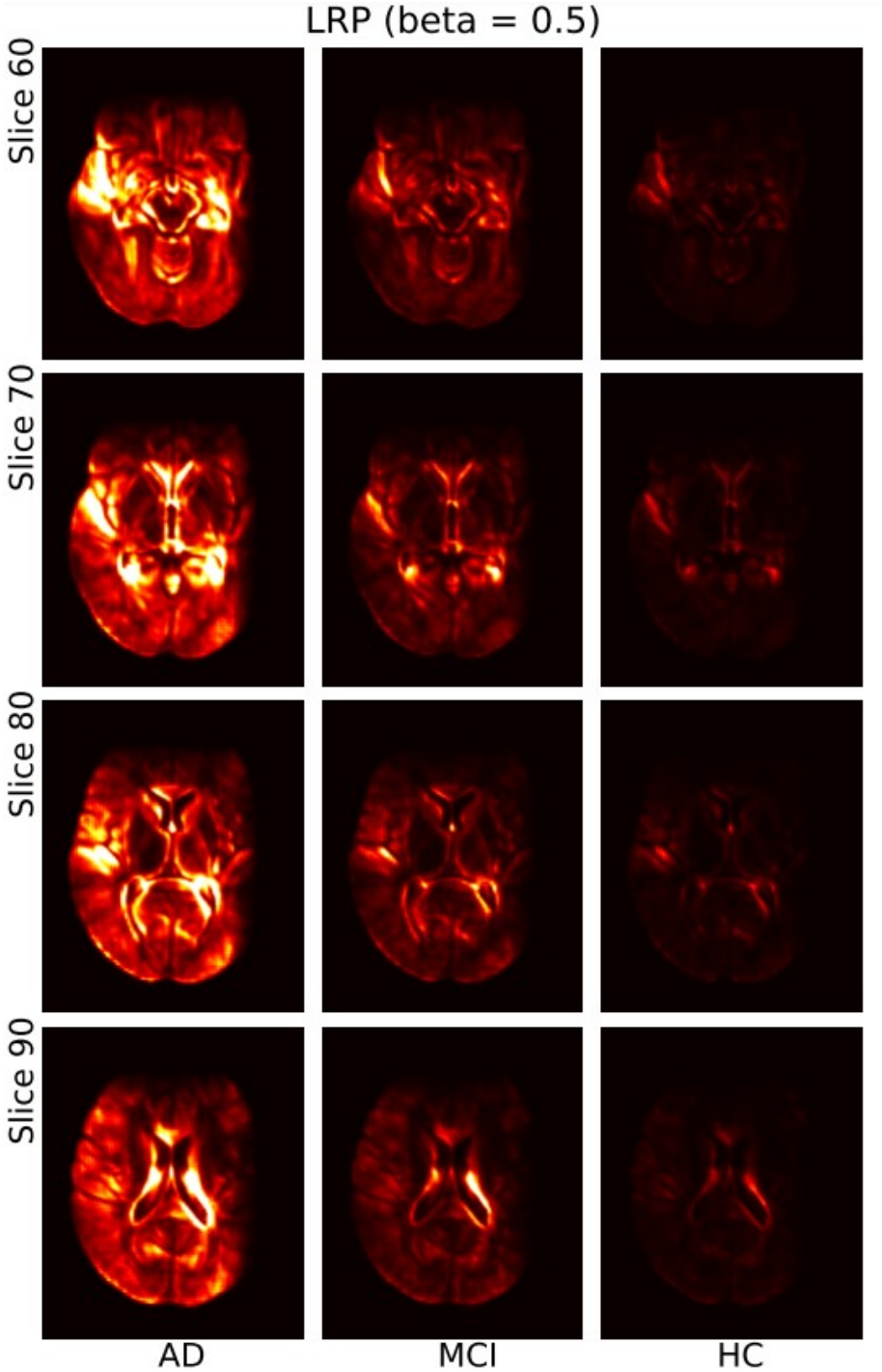

Figure 13: Average heatmaps for AD patients, MCI, and healthy controls (HCs) in the test set are shown for LRP with = 0.5 in multi-class model.The values in the average heatmaps that are higher than the 50th percentile and lower than the 99.5th percentile are linearly color-coded as shown on the scale. Values below (above) these numbers are black (white).

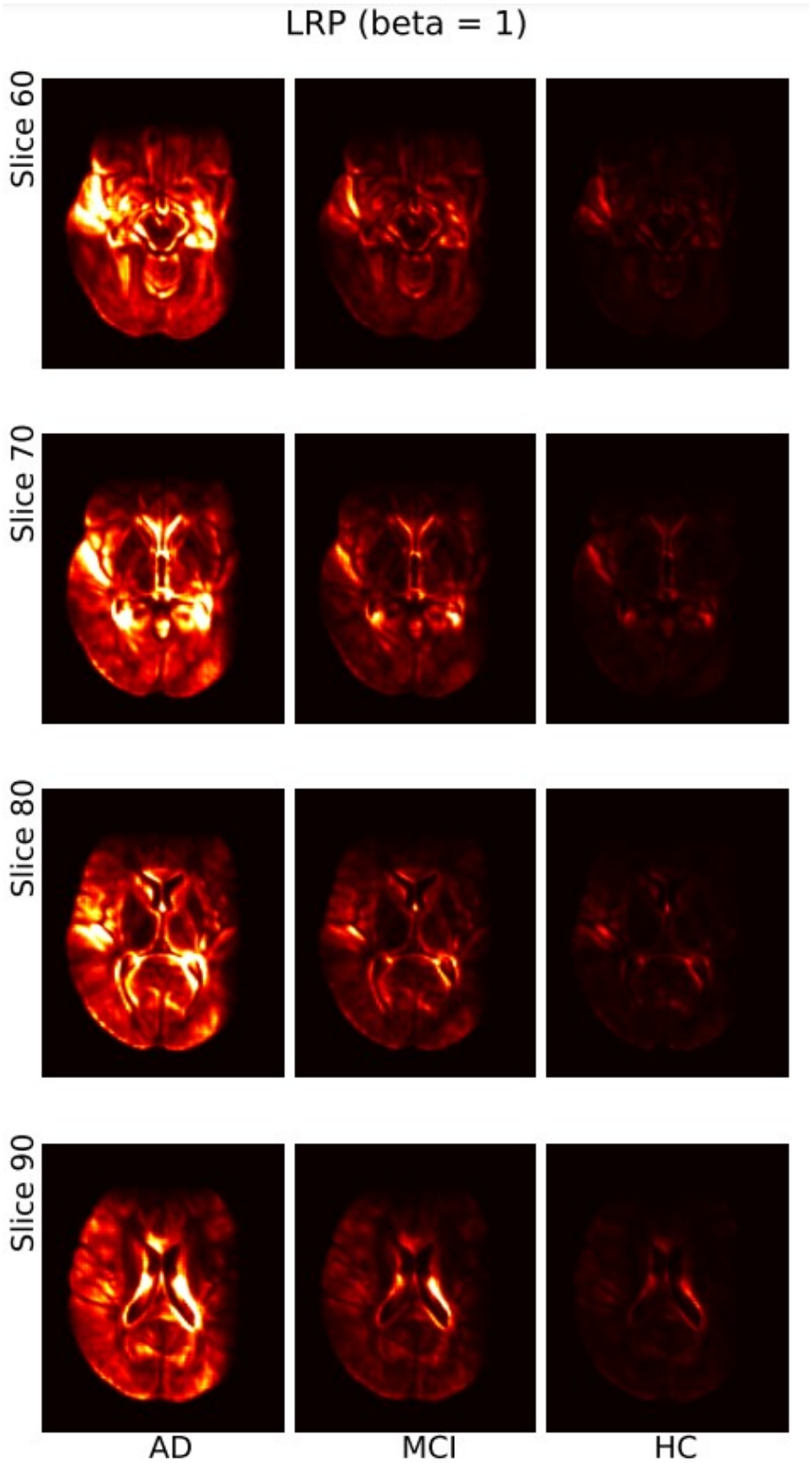

Figure 14: Average heatmaps for AD patients, MCI and healthy controls (HCs) in the test set are shown for LRP with = 1 in multi-class model.The values in the average heatmaps that are higher than the 50th percentile and lower than the 99.5th percentile are linearly color-coded as shown on the scale. Values below (above) these numbers are black (white).

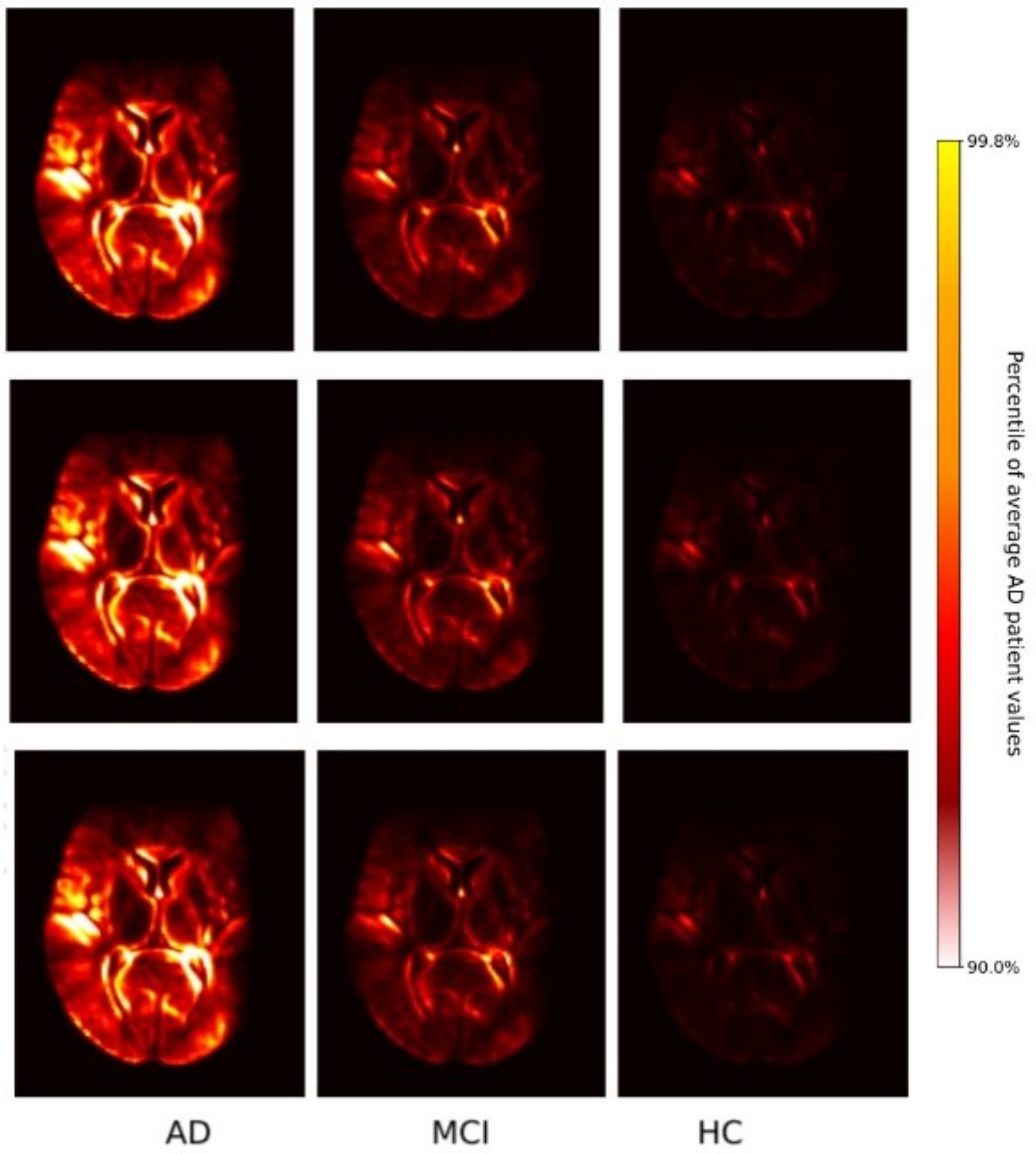

Figure 15: Average heatmaps for AD patients, MCI and healthy controls (HCs) for slice 80 are shown seperately for LRP with = 0 (top slice), 0.5(Middle slice), 1 (bottom slice) in multi-class model. The values in the average heatmaps that are higher than the 90th percentile and lower than the 99.8th percentile are linearly color-coded as shown on the scale. Values below (above) these numbers are black (white).

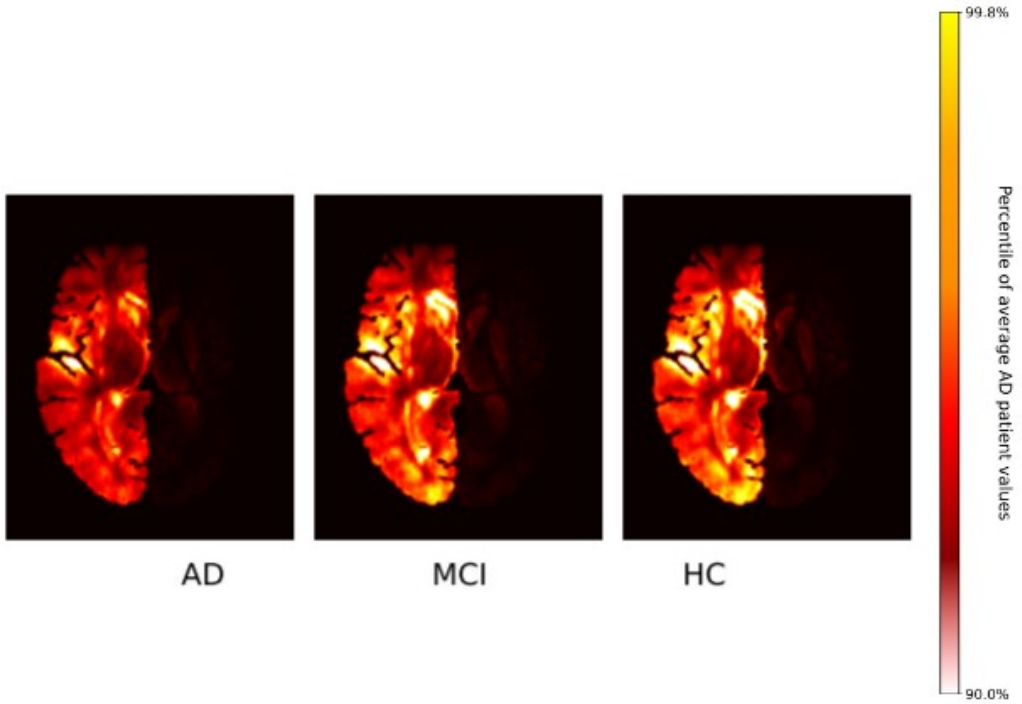

Figure 16: Average heatmaps for AD patients, MCI, and healthy controls (HCs) for slice 80 are shown seperately for GB in multi-class model. The values in the average heatmaps that are higher than the 90th percentile and lower than the 99.8th percentile are linearly color-coded as shown on the scale. Values below (above) these numbers are black (white).

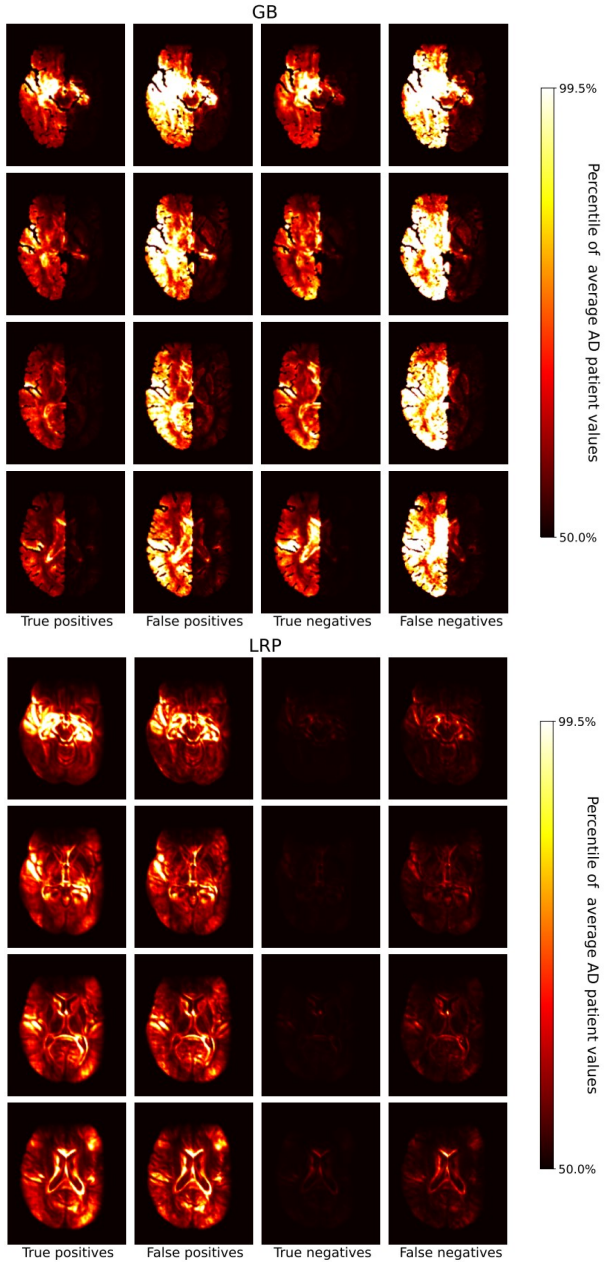

Figure 17: The average heatmaps over all subjects in the test set for binary model are plotted for the following cases (left to right): true positives, false positives, true negatives, and false negatives; separately for LRP with = 0 (bottom) and GB (top). The values smaller than the 50th percentile of the average AD patient in black, increasing values going over red to yellow, and all values greater than the 99.5th percentile in white.

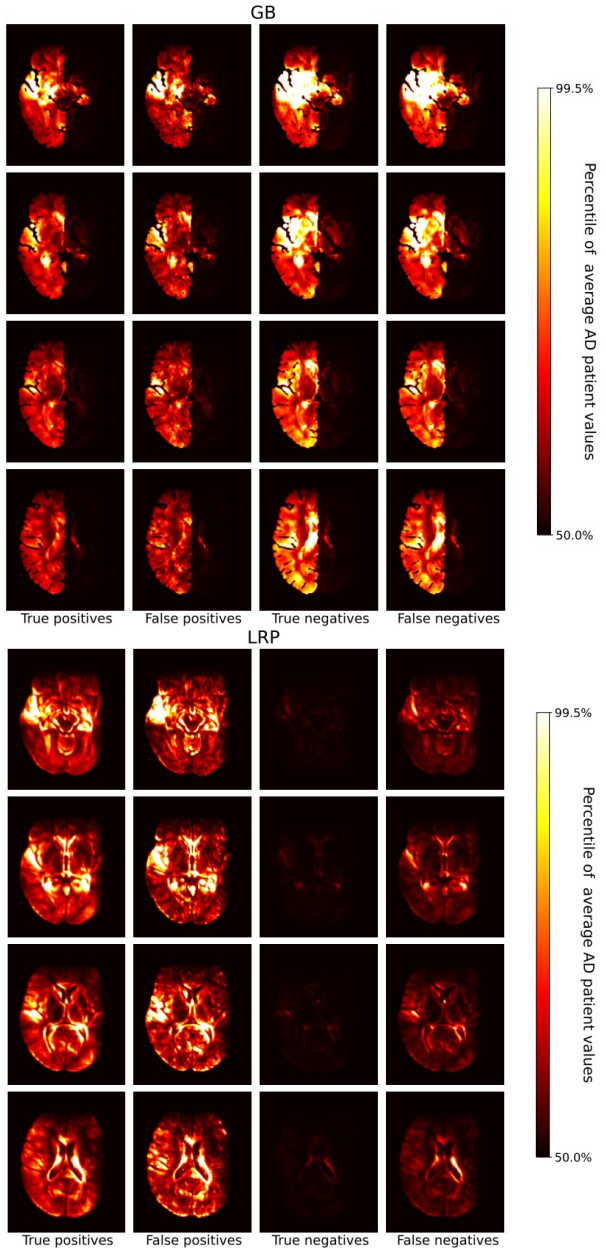

Figure 18: The average heatmaps over all subjects in the test set for multi-class model are plotted for the following cases (left to right): true positives, false positives, true negatives, and false negatives; separately for LRP with = 0 (bottom) and GB (top). The values smaller than the 50th percentile of the average AD patient in black, increasing values going over red to yellow, and all values greater than the 99.5th percentile in white.

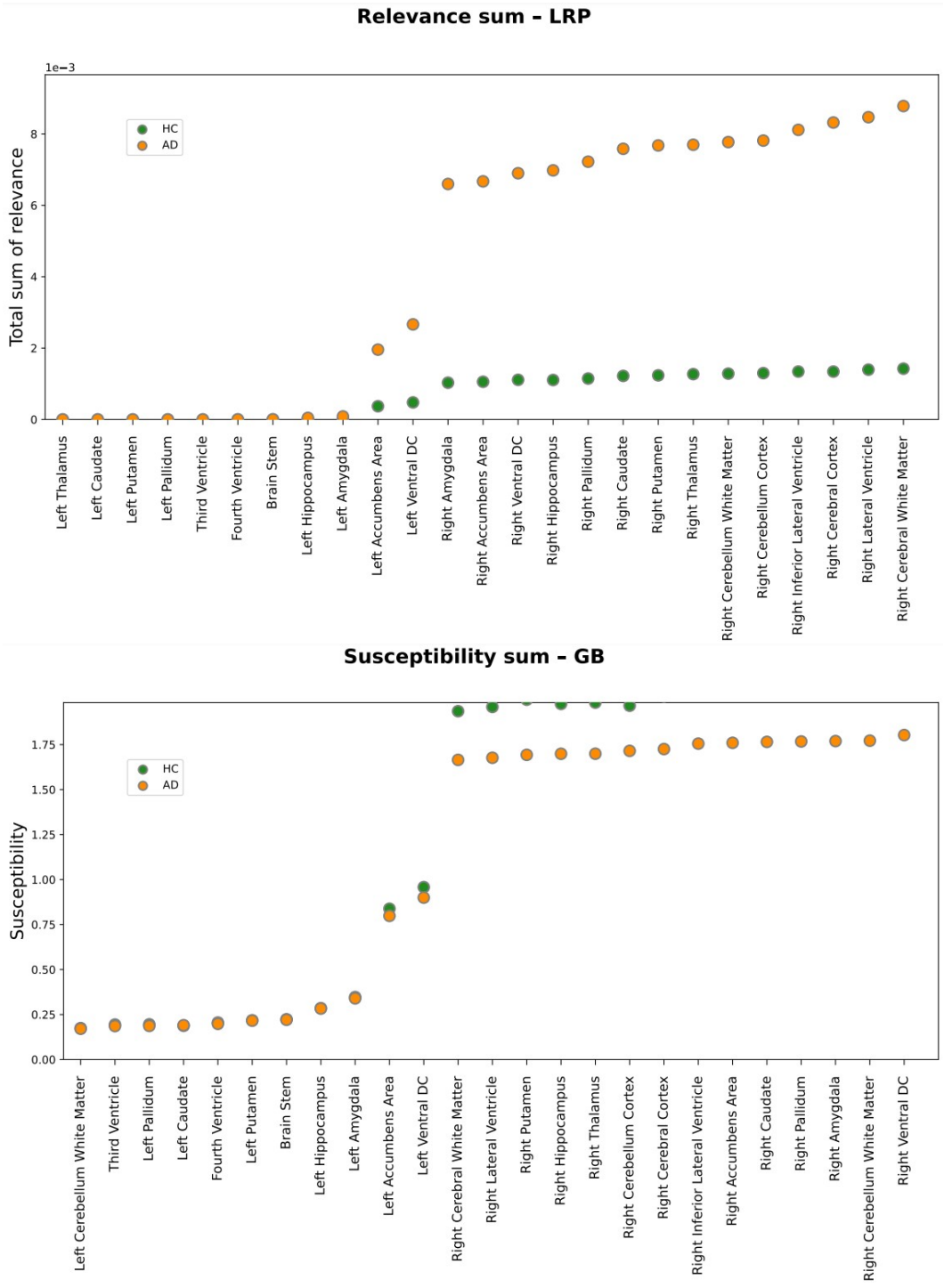

Figure 19: Absolute sum of relevance (LRP, top) and absolute sum of susceptibility (GB, bottom) for binary model is shown for different brain areas. Susceptibility refers to the absolute value of the GB gradients. Only the top 20 most important areas under this metric are shown for LRP and GB respectively. The circles show the average sum for each area over all AD patients (orange) and all healthy controls (HCs, green) in the test set.

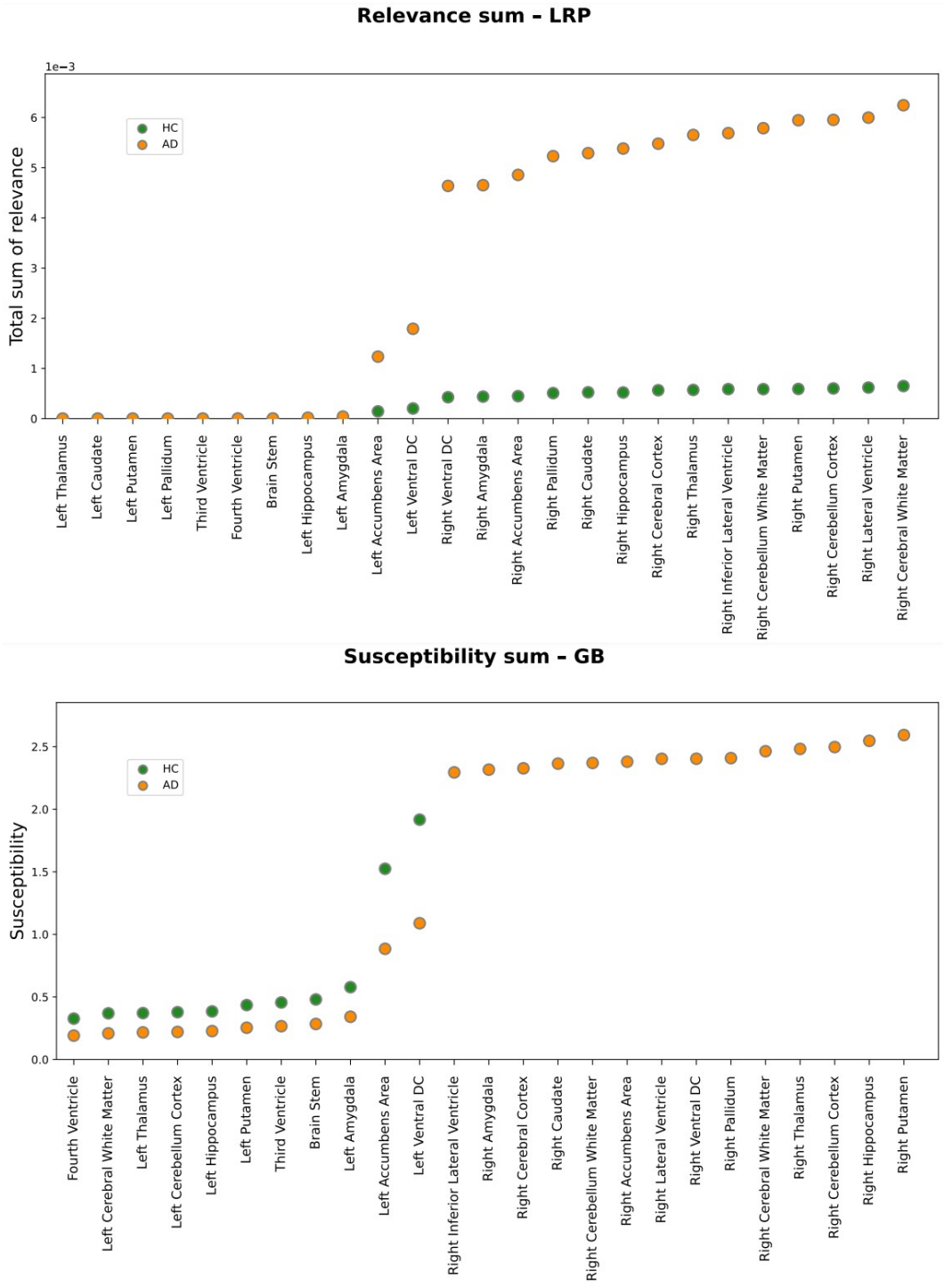

Figure 20: Absolute sum of relevance (LRP, top) and absolute sum of susceptibility (GB, bottom) for multi-class model is shown for different brain areas. Susceptibility refers to the absolute value of the GB gradients. Only the top 20 most important areas under this metric are shown for LRP and GB respectively. The circles show the average sum for each area over all AD patients (orange) and all healthy controls (HCs, green) in the test set.

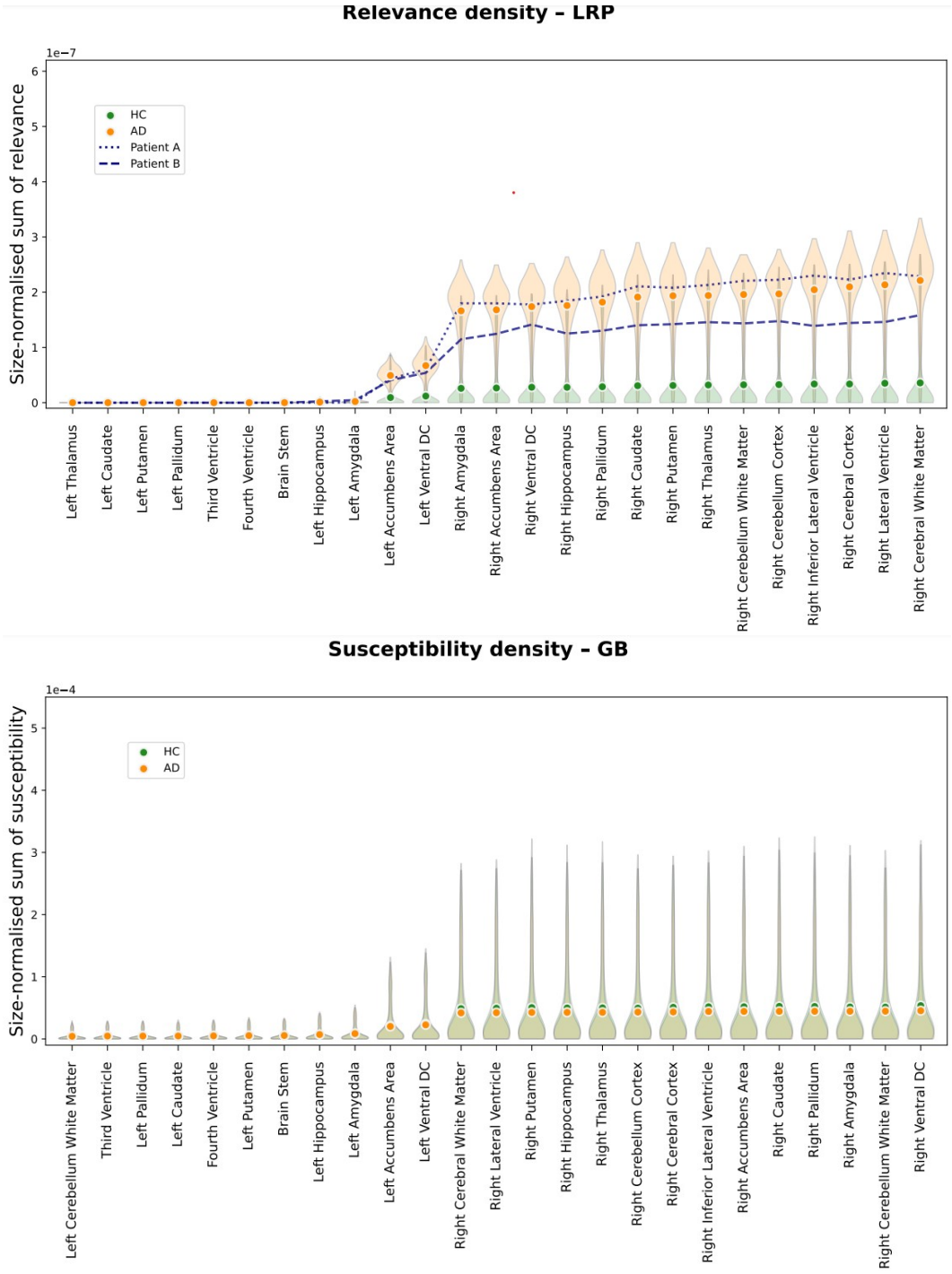

Figure 21: Size-normalized relevance (LRP, top) and size-normalized susceptibility (GB, bottom) is shown for different brain areas. Only the top 16 most important areas under this metric are shown for LRP and GB respectively for binary model.

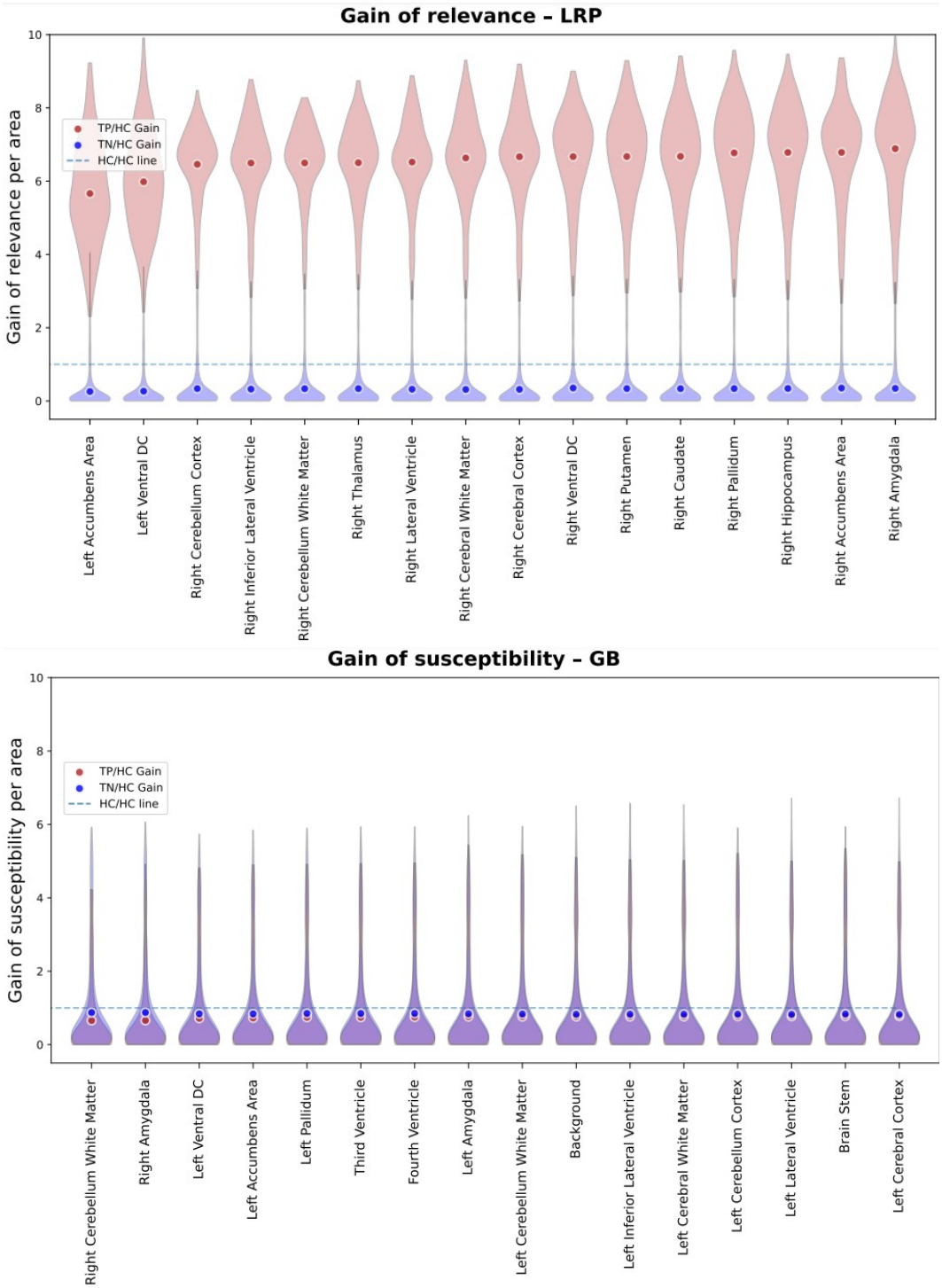

Figure 22: Gain ofrelevance (LRP, top) and gain of susceptibility (GB, bottom) for binary classification model.

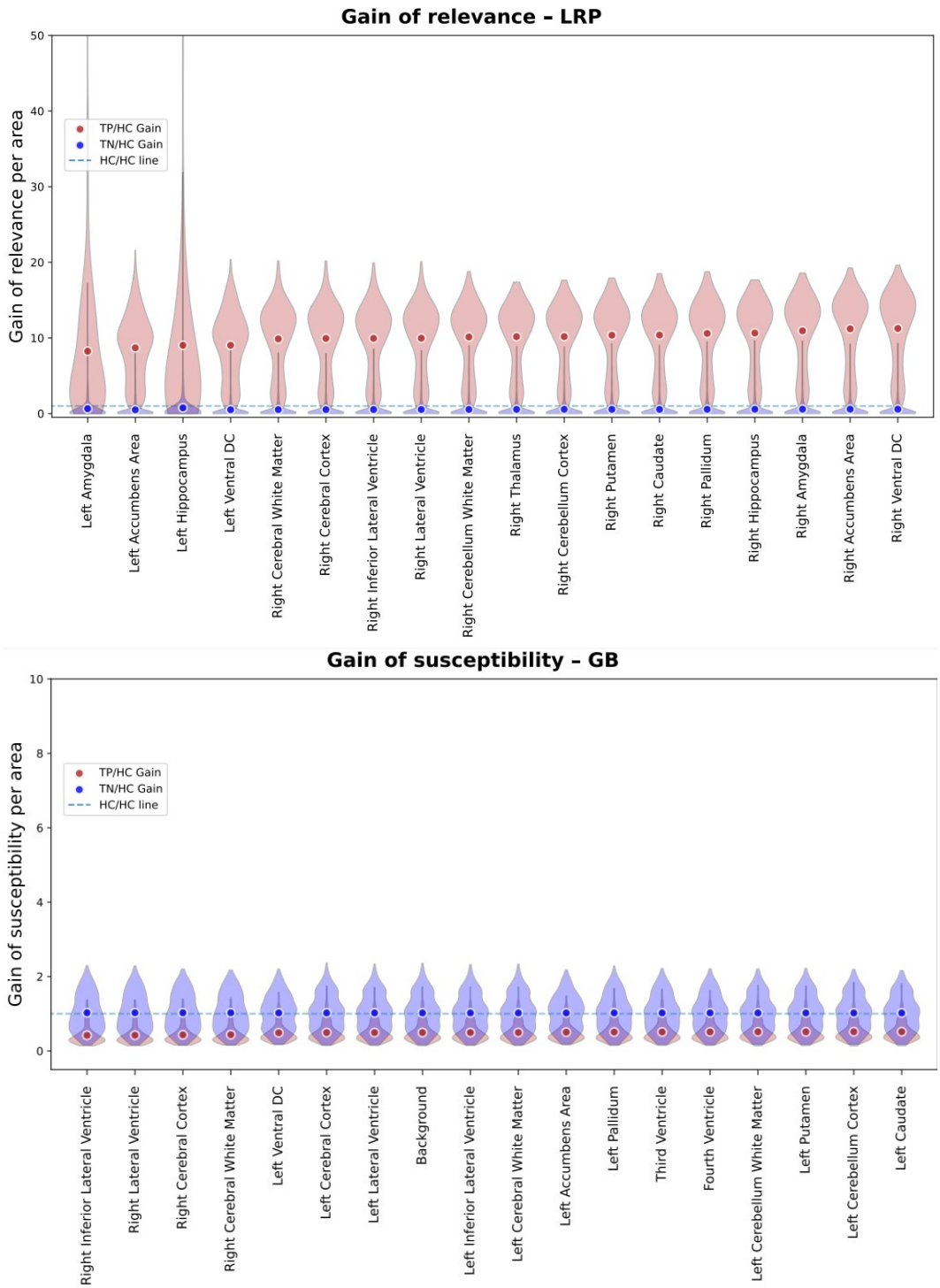

Figure 23: Gain ofrelevance (LRP, top) and gain of susceptibility (GB, bottom) for multi-class classification model.

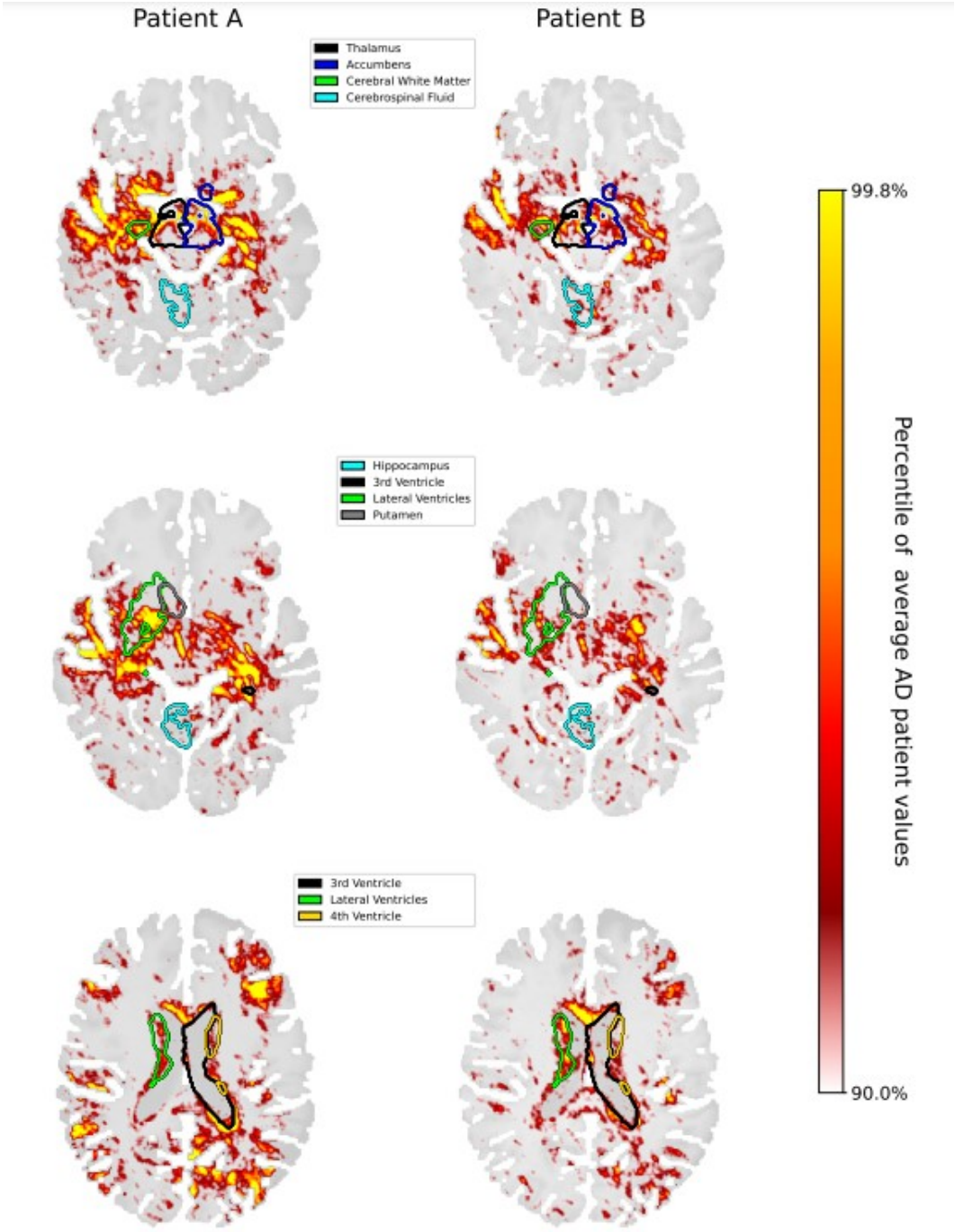

Figure 24: Three brain slices are shown for patient A and patient B for binary model when beta is 0. The highlighted areas are the Thalamus, Accumbens, CWM, CSF Hippocampus, 3rd Ventricle, Lateral Ventricles, Putamen, 3rd Ventricle, Lateral Ventricles, 4th Ventricle. The scale for the heatmap is chosen relative to the average AD patient heatmap. Hence, values in the individual patients that are higher than the 90th percentile and lower than the 99.5th percentile are linearly color-coded as shown on the scale. Values below (above) these numbers are transparent (yellow).

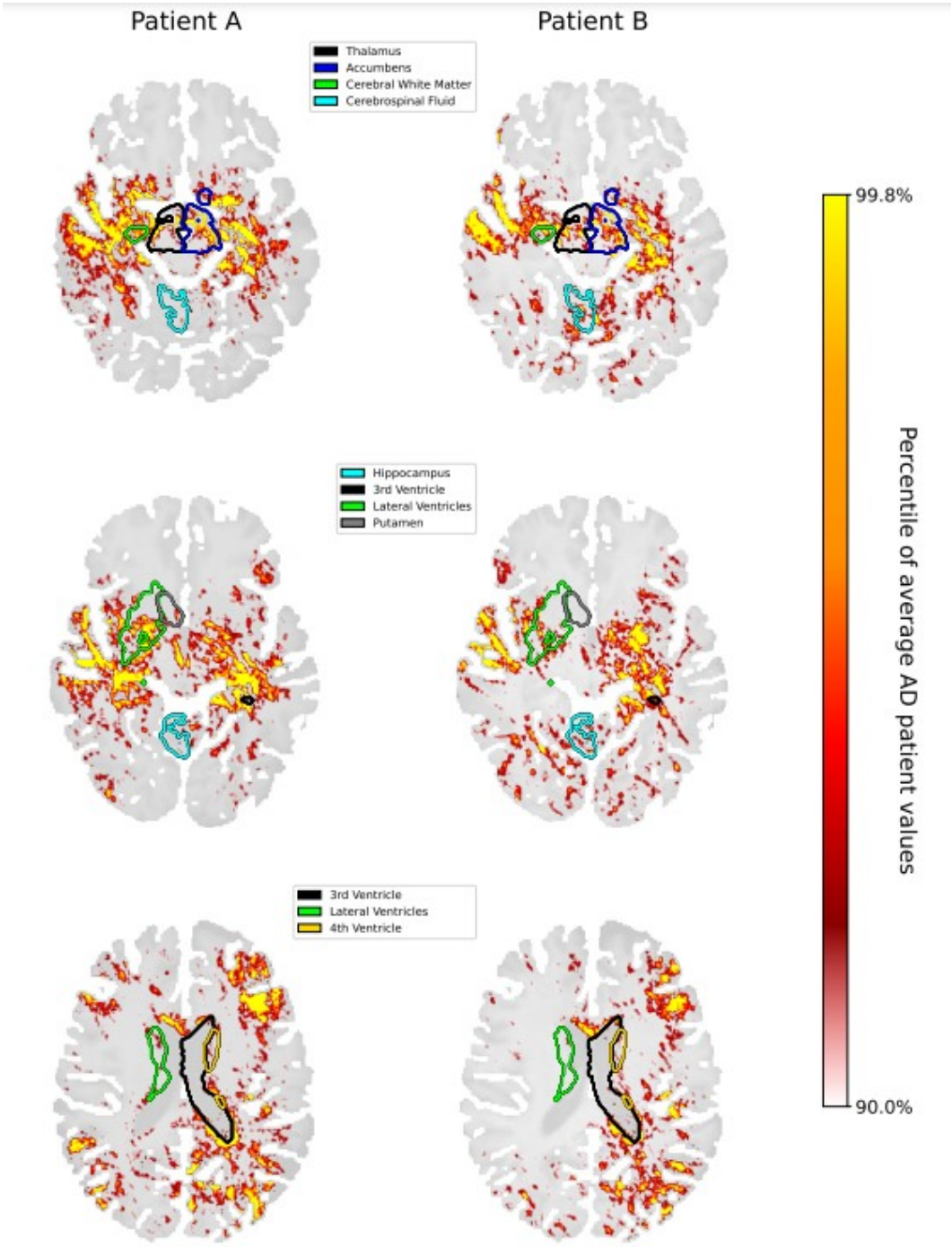

Figure 25: Three brain slices are shown for patient A and patient B for binary model when beta is 0.5. The highlighted areas are the Thalamus, Accumbens, CWM, CSF Hippocampus, 3rd Ventricle, Lateral Ventricles, Putamen, 3rd Ventricle, Lateral Ventricles, 4th Ventricle. The scale for the heatmap is chosen relative to the average AD patient heatmap. Hence, values in the individual patients that are higher than the 90th percentile and lower than the 99.5th percentile are linearly color-coded as shown on the scale. Values below (above) these numbers are transparent (yellow)

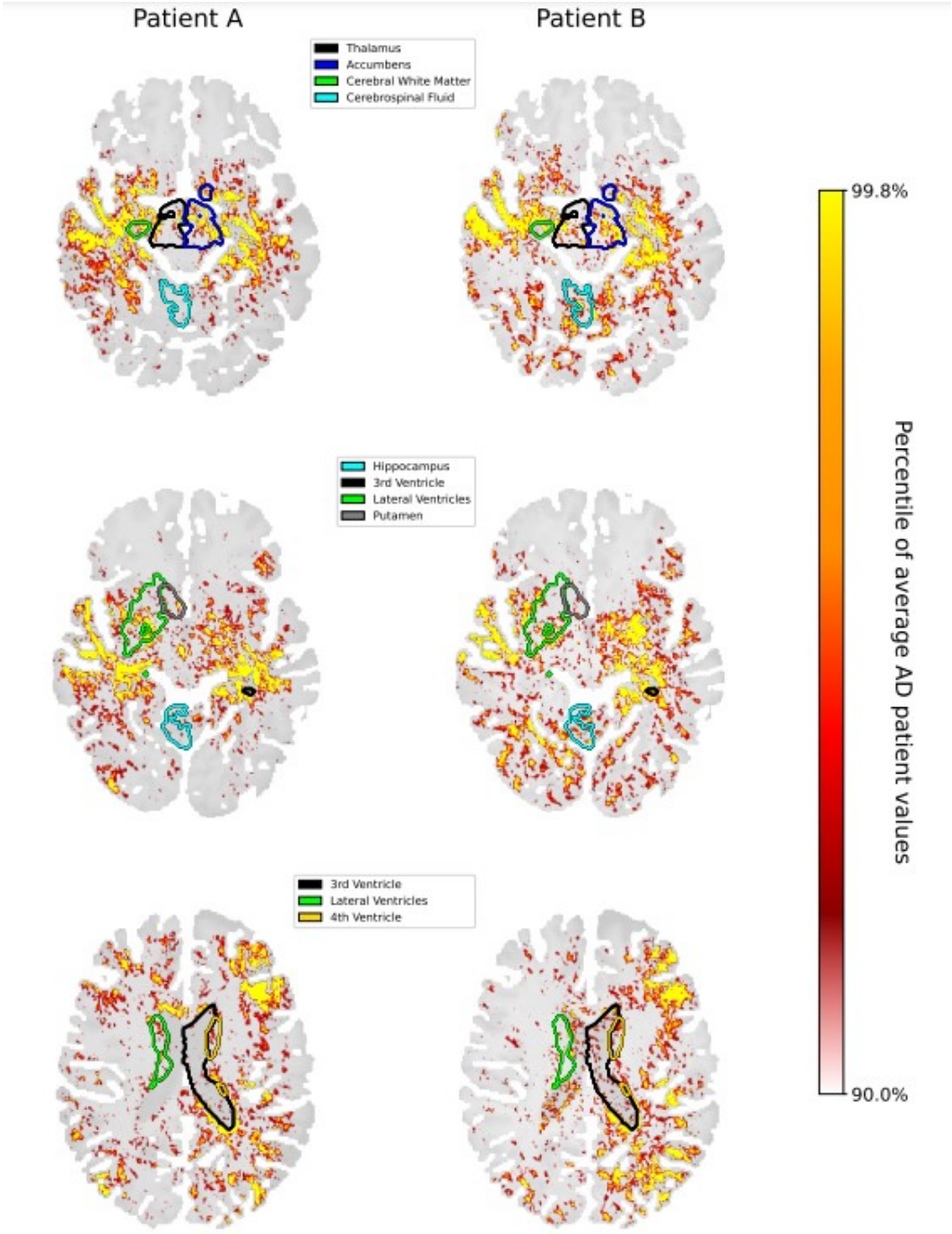

Figure 26: Three brain slices are shown for patient A and patient B for binary model when beta is 1. The highlighted areas are the Thalamus, Accumbens, CWM, CSF Hippocampus, 3rd Ventricle, Lateral Ventricles, Putamen, 3rd Ventricle, Lateral Ventricles, 4th Ventricle. The scale for the heatmap is chosen relative to the average AD patient heatmap. Hence, values in the individual patients that are higher than the 90th percentile and lower than the 99.5th percentile are linearly color-coded as shown on the scale. Values below (above) these numbers are transparent (yellow)

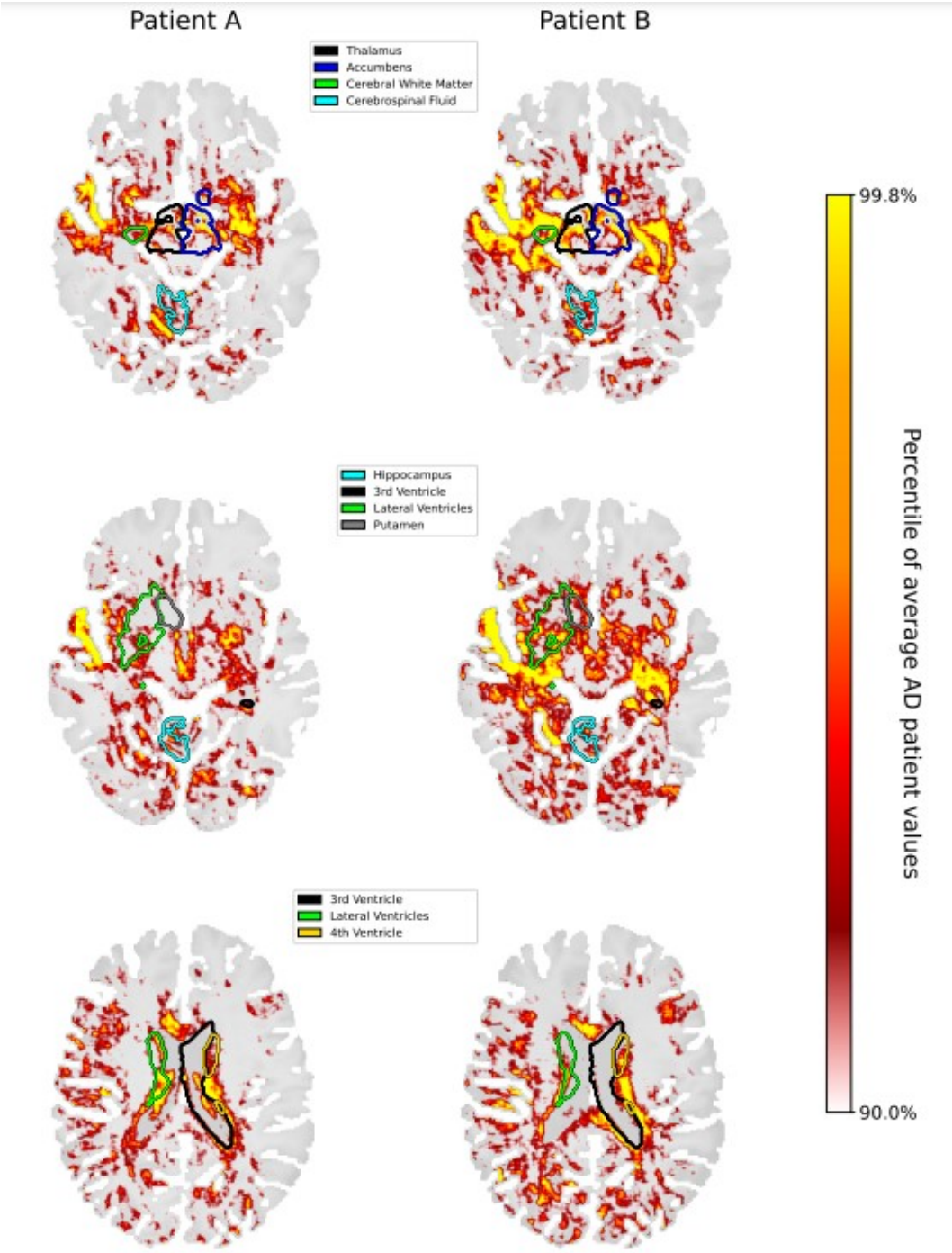

Figure 27: Three brain slices are shown for patient A and patient B for multi-class model when beta is 0. The highlighted areas are the Thalamus, Accumbens, CWM, CSF Hippocampus, 3rd Ventricle, Lateral Ventricles, Putamen, 3rd Ventricle, Lateral Ventricles, 4th Ventricle. The scale for the heatmap is chosen relative to the average AD patient heatmap. Hence, values in the individual patients that are higher than the 90th percentile and lower than the 99.5th percentile are linearly color-coded as shown on the scale. Values below (above) these numbers are transparent (yellow)

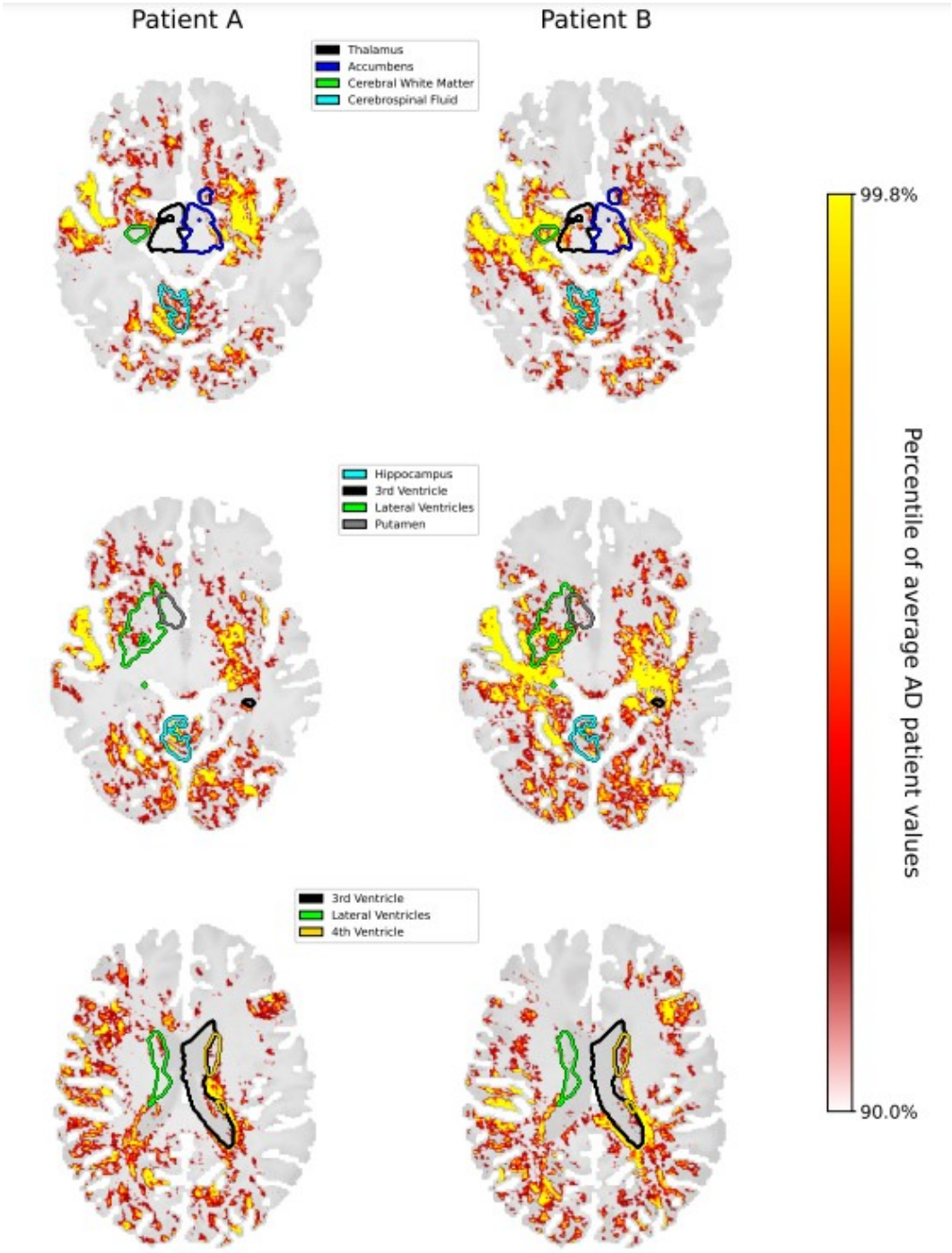

Figure 28: Three brain slices are shown for patient A and patient B for multi-class model when beta is 0.5. The highlighted areas are the Thalamus, Accumbens, CWM, CSF Hippocampus, 3rd Ventricle, Lateral Ventricles, Putamen, 3rd Ventricle, Lateral Ventricles, 4th Ventricle. The scale for the heatmap is chosen relative to the average AD patient heatmap. Hence, values in the individual patients that are higher than the 90th percentile and lower than the 99.5th percentile are linearly color-coded as shown on the scale. Values below (above) these numbers are transparent (yellow)

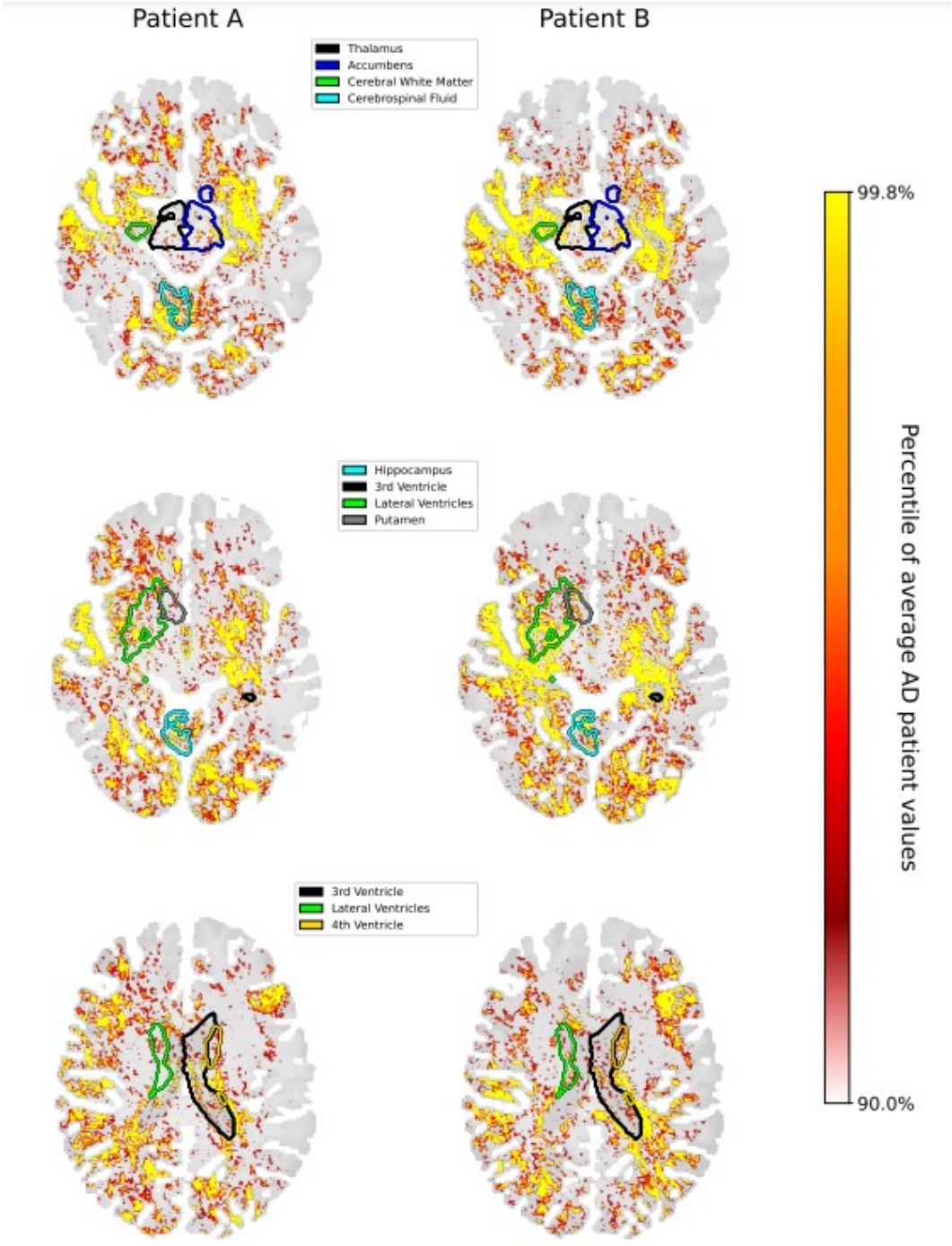

Figure 29: Three brain slices are shown for patient A and patient B for multi-class model when beta is 1. The highlighted areas are the Thalamus, Accumbens, CWM, CSF Hippocampus, 3rd Ventricle, Lateral Ventricles, Putamen, 3rd Ventricle, Lateral Ventricles, 4th Ventricle. The scale for the heatmap is chosen relative to the average AD patient heatmap. Hence, values in the individual patients that are higher than the 90th percentile and lower than the 99.5th percentile are linearly color-coded as shown on the scale. Values below (above) these numbers are transparent (yellow)