Layer-Wise Relevance Propagation for Explaining Deep Neural Network Decisions in MRI-Based Alzheimer's Disease Classification

Moritz Böhle[1,2‡], Fabian Eitel[1,2‡], Martin Weygandt[1,3] and Kerstin Ritter[1,2*]
on behalf of the Alzheimer's Disease Neuroimaging Initiative[†]

[1] Berlin Institute of Health, Charité—Universitätsmedizin Berlin, Corporate Member of Freie Universität Berlin, Humboldt-Universität zu Berlin, Berlin, Germany, [2] Department of Psychiatry and Psychotherapy, Bernstein Center for Computational Neuroscience, Berlin, Germany, [3] Excellence Cluster NeuroCure Berlin, Berlin, Germany

Deep neural networks have led to state-of-the-art results in many medical imaging tasks including Alzheimer's disease (AD) detection based on structural magnetic resonance imaging (MRI) data. However, the network decisions are often perceived as being highly non-transparent, making it difficult to apply these algorithms in clinical routine. In this study, we propose using layer-wise relevance propagation (LRP) to visualize convolutional neural network decisions for AD based on MRI data. Similarly to other visualization methods, LRP produces a heatmap in the input space indicating the importance/relevance of each voxel contributing to the final classification outcome. In contrast to susceptibility maps produced by guided backpropagation ("Which change in voxels would change the outcome most?"), the LRP method is able to directly highlight positive contributions to the network classification in the input space. In particular, we show that (1) the LRP method is very specific for individuals ("Why does this person have AD?") with high inter-patient variability, (2) there is very little relevance for AD in healthy controls and (3) areas that exhibit a lot of relevance correlate well with what is known from literature. To quantify the latter, we compute size-corrected metrics of the summed relevance per brain area, e.g., relevance density or relevance gain. Although these metrics produce very individual "fingerprints" of relevance patterns for AD patients, a lot of importance is put on areas in the temporal lobe including the hippocampus. After discussing several limitations such as sensitivity toward the underlying model and computation parameters, we conclude that LRP might have a high potential to assist clinicians in explaining neural network decisions for diagnosing AD (and potentially other diseases) based on structural MRI data.

Keywords: Alzheimer's disease, MRI, visualization, explainability, layer-wise relevance propagation, deep learning, convolutional neural networks (CNN)

1. INTRODUCTION

In the 2018 World Alzheimer Report, it was estimated that 50 million people worldwide were suffering from dementia and this number was projected to rise to more than 152 million people until 2050. The most common reason for dementia is Alzheimer's disease (AD) accounting for around 60–70 % of dementia cases (WHO, 2017). AD is characterized by abnormal cell death, primarily in the medial temporal lobe. This cell death is thought to be rooted in protein plaques and neurofibrillary tangles, which restrict normal neural function (Bondi et al., 2017). The resulting atrophy is visible in structural magnetic resonance imaging (MRI) data, and derived markers (such as hippocampal volume or gray matter density) have been used to diagnose AD and predict disease progression (Frisoni et al., 2010; Rathore et al., 2017). In the last decade, those markers have frequently been employed in machine learning settings to allow for predictions on an individual level (Klöppel et al., 2008; Orrù et al., 2012; Weiner et al., 2013; Ritter et al., 2015, 2016). However, those expert features usually reflect only one part of disease pathology and the combination with standard machine learning methods, such as support vector machines, do not allow for finding new and potentially unexpected hidden data characteristics that might also be important to describe a disease.
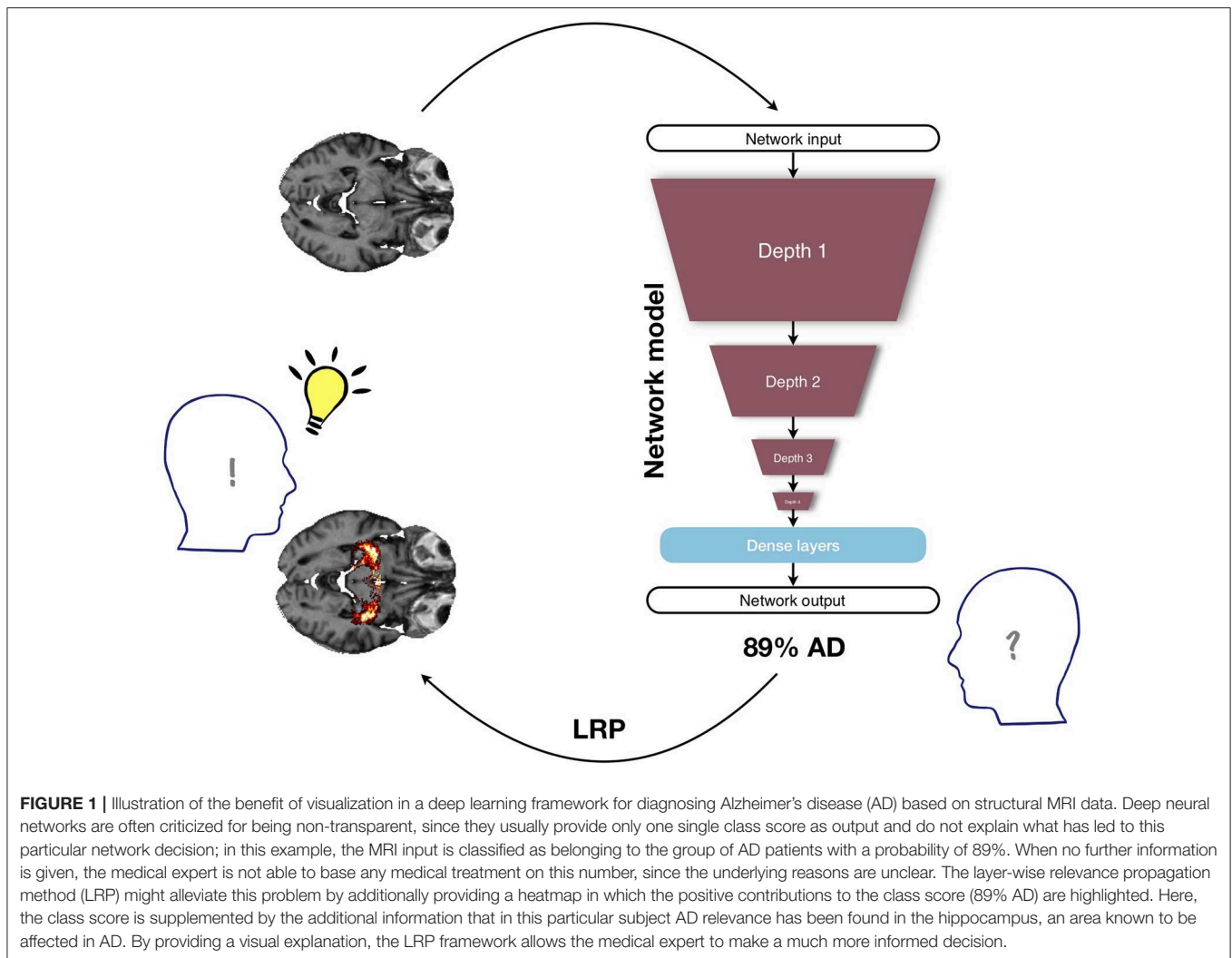
By extracting hierarchical information directly from raw or minimally processed data, deep learning approaches (LeCun et al., 2015) can help to fill a gap here and offer a great potential for improving automatic disease diagnostics. One family of algorithms that perfectly lends itself to perform non-linear feature extraction from image data and their respective classification into disease categories are convolutional neural networks (CNNs), a type of (deep) neural networks optimized for image data. The key idea behind CNNs is inspired by the mechanism of receptive fields in the primate's visual cortex and relates to the application of local convolutional filters for extracting regional information (LeCun and Bengio, 1995). They typically consist of a sequence of convolutional and pooling layers which allow for summarizing key characteristics into feature maps. These feature maps can then be used by a fully-connected layer or any other classifier for solving the primary supervised learning problem (e.g., AD classification). CNNs have been proven to be very successful in a wide range of medical imaging applications (Litjens et al., 2017), including AD detection based on neuroimaging data (e.g., Gupta et al., 2013; Suk et al., 2014; Payan and Montana, 2015; Sarraf and Tofighi, 2016; Korolev et al., 2017; for a review, see Vieira et al., 2017).

Despite this success, automatically learning the features comes at a cost: the decisions of neural networks are notoriously hard to interpret in retrospect. Therefore, deep learning methods, including CNNs, often face the criticism that they are "black-box" (Castelvecchi, 2016). In contrast to some simpler learning algorithms, in particular decision trees, they do not offer a simple and comprehensible explanation; their architecture is complex and consists of several to many layers with hundreds of thousands parameters that need to be trained. In the medical domain, however, it is imperative to base diagnoses and subsequent

treatments on an informed decision and not on a single yes/no answer of an algorithm. Therefore, if CNNs should support clinicians in their daily work, ways have to be found to visualize and interpret the network's "decision" (see **Figure 1**). In the last years, a number of suggestions have been made to visualize *what* is actually learned by a CNN. Besides straightforward methods such as the extraction of activations during convolution or the visualization of weights, among the most well-known techniques for visualization are the sensitivity analysis 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 (2014). In Alzheimer's research only a very few studies exist that looked into such visualization methods (Esmaeilzadeh et al., 2018; Rieke et al., 2018; Yang et al., 2018).

Most promising for the use in the medical imaging domain is the generation of an individual heatmap for each patient, which lies in the same space as the input image and indicates the importance of each voxel for the final (individual) classification decision. By allowing for a human-guided, intuitive investigation of what drives the classifier to come to a certain classification decision, individual heatmaps hold great potential in assisting and understanding diagnostic decisions performed by deep neural networks. However, for any visualization method that produces heatmaps, it is very important to understand how they are computed and what their limitations are. In natural images, for example, it has been argued that methods relying on gradients (e.g., sensitivity analysis or guided backpropagation) only measure the susceptibility of the output to changes in the input and might not necessarily coincide with those areas on which the network bases its decision. A powerful method to overcome this limitation is layer-wise relevance propagation (LRP, Bach et al., 2015), which decomposes the network's output score (e.g., for AD) into the individual contributions of the input neurons while keeping the total amount of relevance constant across layers (conservation principle). In contrast to showing "susceptibility maps" as gradient-based methods, the heatmap does not rely on gradients, but takes into account model parameters (i.e., weights) and neuron activations (Bach et al., 2015; Samek et al., 2015). By this, the heatmaps are less prone to group effects in the data. Intuitively, LRP has the potential to answer the question "what speaks for AD in this particular patient?" as opposed to "which change in voxels would change the outcome most?" addressed in gradient-based approaches. In terms of explainability, LRP has been shown to be superior to those gradient methods and deconvolution methods in three natural imaging data sets (Samek et al., 2015). In cognitive neuroscience, the LRP method has been recently applied to single-trial EEG and functional MRI classification (Sturm et al., 2016; Thomas et al., 2018). To the best of our knowledge, it has so far not been applied in clinical disease classification based on structural MRI data.

In this study, we use LRP to explain individual classification decisions for AD patients and healthy controls (HCs) based on a CNN trained on structural MRI data (T1-weighted MPRAGE) from the Alzheimer's Disease Neuroimaging Initiative

FIGURE 1 | Illustration of the benefit of visualization in a deep learning framework for diagnosing Alzheimer's disease (AD) based on structural MRI data. Deep neural networks are often criticized for being non-transparent, since they usually provide only one single class score as output and do not explain what has led to this particular network decision; in this example, the MRI input is classified as belonging to the group of AD patients with a probability of 89%. When no further information is given, the medical expert is not able to base any medical treatment on this number, since the underlying reasons are unclear. The layer-wise relevance propagation method (LRP) might alleviate this problem by additionally providing a heatmap in which the positive contributions to the class score (89% AD) are highlighted. Here, the class score is supplemented by the additional information that in this particular subject AD relevance has been found in the hippocampus, an area known to be affected in AD. By providing a visual explanation, the LRP framework allows the medical expert to make a much more informed decision.

(ADNI[1]). Based on the trained CNN model, we generated LRP heatmaps for each subject in the test set. Importantly, each heatmap indicates the voxel-wise relevance for the particular classification decision (AD or HC). To spot the most relevant regions for AD classification, we computed average heatmaps across AD patients and HCs, which we then further split into correct and wrong classification decisions (i.e., true positives, false positives, true negatives, false negatives). To analyze the relevance in different brain areas according to the Scalable Brain Atlas by Neuromorphometrics Inc. (Bakker et al., 2015), we suggest size-corrected metrics and compared these metrics between LRP and guided backpropagation. We have chosen guided backpropagation as a baseline method because (1) sensitivity analysis is the most common method for generating heatmaps, (2) it results in more focused heatmaps compared to only using backpropagation (Rieke et al., 2018) and (3) it is better comparable to LRP than occlusion methods with respect to our relevance measures. On an individual level, we analyzed the

heatmap patterns of single subjects ("relevance fingerprinting") and correlate them with the hippocampal volume as a key biomarker of AD. We show that the LRP heatmaps succeeded in depicting individual contributions to AD diagnosis and might hold great potential as a diagnostic tool.

2. MATERIALS AND METHODS

2.1. Data and Preprocessing

Data used in the preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI, RRID:SCR_003007) database (adni.loni.usc.edu). The ADNI was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimer's disease (AD). For up-to-date information, see www.adni-info.org.

[1]http://adni.loni.usc.edu/

We included structural MRI data of all subjects with Alzheimer's disease (AD) and healthy controls (HCs) listed in the "MRI collection - Standardized 1.5T List - Annual 2 year". The subjects in the data set are labeled as AD if the Clinical Dementia Rating (CDR) score (Morris, 1993) was greater than 0.5. HCs are selected as those subjects with a CDR score of 0. In total, we included 969 individual scans (475 AD, 494 HC) of 193 AD patients and 151 HCs (up to three time points). All scans were acquired with 1.5 T scanners at various sites and had undergone gradient non-linearity, intensity inhomogeneity and phantom-based distortion correction. We downloaded T1-weighted MPRAGE scans and non-linearly registered them to the 1mm resolution 2009c version of the ICBM152 reference brain using Advanced Normalization Tools (ANTs[2]). This has been done to (1) ensure a relative alignment across subjects, (2) allow the convolutional neural network to extract more robust features, and (3) be able to analyze the heatmaps in a common space. For the region-wise analysis of heatmaps, we used the Scalable Brain Atlas by Neuromorphometrics Inc. (Bakker et al., 2015) available in SPM12[3]. A list of all areas included can be found in the SPM12 package.

2.2. Convolutional Neural Network Architecture

Convolutional neural networks (CNNs) are neural networks optimized for array data including images or videos (LeCun et al., 2015). In addition to input and output layer, they consist of several hidden layers including convolutional and pooling layers. In convolutional layers, in contrast to fully-connected layers, the weights and the bias terms are shared between all neurons in a given layer for a given filter. This means that each of the neurons applies the same *filter* or *kernel* to the input, but at a different position, usually with a displacement (often called stride) of 1–3 between neighboring neurons. Since these filters are learned via the backpropagation algorithm, CNNs do not rely on hand-crafted features, but can be applied to minimally processed data (LeCun et al., 2015). CNNs have been very successfully applied to a large number of applications including image and speech recognition (Krizhevsky et al., 2012; Abdel-Hamid et al., 2014; Long et al., 2015) as well as medical imaging and AD classification based on MRI data (Gupta et al., 2013; Suk et al., 2014; Payan and Montana, 2015; Sarraf and Tofighi, 2016; Korolev et al., 2017; Litjens et al., 2017; Vieira et al., 2017).

The model in the present study consists of four convolutional blocks followed by two fully-connected layers. Each block features a convolutional layer with f filters ($f = 8, 16, 32, 64$) and filter sizes of $3 \times 3 \times 3$. Every convolutional layer is followed by batch normalization and max pooling with window sizes $w \times w \times w$ ($w = 2, 3, 2, 3$). The fully-connected layers contain 128 and 2 units respectively and dropout ($p = 40\%$) is applied before each. The final fully-connected layer, which is activated by a softmax function serves as the network output, providing the class scores for HCs (first unit) and AD (second unit) respectively. As an optimizer Adam (Kingma and Ba, 2015) was used with an

initial learning rate of 0.0001 and a weight decay of 0.0001. The data was split into a training data set (163 AD patients, 121 HCs; 797 images in total), a validation set for optimizing the hyperparameters (18 AD patients, 18 HCs; 100 images in total) and a test set (30 AD patients, 30 HCs; 172 images in total). To ensure independence between training and test data, we performed the split of the data on the level of patients instead of images. The data was augmented during training by flipping the images along the sagittal axis ($p = 50\%$) and translated along the sagittal axis between –2 and 2 voxels. When the model did not improve for 8 epochs on the validation set, training was stopped. The training epoch (i.e., model checkpoint) with the best validation accuracy (91.00%) was then applied to the test data, resulting in a classification accuracy of 87.96%.

2.3. Visualization Methods
2.3.1. Layer-Wise Relevance Propagation (LRP)
In the following, we will introduce the Layer-wise Relevance Propagation (LRP) algorithm by Bach et al. (2015). The core idea underlying the LRP algorithm for attributing relevance to individual input nodes is to trace back contributions to the final output node layer by layer. While several different versions of the LRP algorithm exist, they all share the same principle: the total relevance—e.g., the activation strength of an output node for a certain class—is conserved per layer; each of the nodes in layer l that contributed to the activation of a node j in the subsequent layer $l + 1$ gets attributed a certain share of the relevance R_{l+1}^{j} of that node. Overall, the sum over the relevances of all nodes i contributing to neuron j in layer l must sum to R_{l+1}^{j}, such that the total relevance per layer is conserved:

$$\sum_i R_{l,l+1}^{i \to j} = R_{l+1}^{j} \quad (1)$$

There are different ways in which the relevance can be distributed over the input nodes i and different rules for how to distribute the relevances have been proposed. In this paper, we used the β-rule (as described in Binder et al., 2016b):

$$R_{l,l+1}^{i \to j} = \left((1 + \beta) \frac{z_{ij}^{+}}{z_{j}^{+}} - \beta \frac{z_{ij}^{-}}{z_{j}^{-}} \right) R_{l+1}^{j}. \quad (2)$$

Here, $z_{ij}^{+/-}$ refers to the amount of positive/negative input that node i contributed to node j. The individual contributions are divided by the sum over all positive/negative contributions of the nodes in layer l, $z_{j}^{+/-} = \sum_i z_{ij}^{+/-}$, such that the relevance is conserved from layer $l + 1$ to layer l. We have chosen this rule, as it allows for adjusting how much weight is put on positive contributions relative to inhibitory contributions that benefit the AD score. LRP with a β value of zero allows only positive contributions to be shown in the heatmap, whereas non-zero β values additionally correct for the inhibitory effects of neuron activations. When diagnosing AD, the network needs to balance structural evidence speaking for and against AD. Any given local area that looks *healthy* to the network, might have inhibitory effects on the AD score, as it correlates more with HC patients.

---

[2] http://stnava.github.io/ANTs/
[3] http://www.fil.ion.ucl.ac.uk/spm/software/spm12/

As the network increases its receptive field size throughout the layers, healthy areas within this receptive field might inhibit the contribution of affected areas to the final class score of AD. By reversing this process with LRP, positive contributions lying closer to healthy areas will thus obtain a lower relevance score, as they overlap with inhibited receptive fields. This leads to sparser heatmaps, see also Binder et al. (2016a), and might disproportionately affect small structures surrounded by "healthy areas." As AD—especially in the early stages of the disease—can affect brain areas in a highly localized manner, heatmaps obtained with lower β values might therefore be more meaningful, as they highlight *all* positive contributions, irrespective of their surroundings. Accordingly, we focus in the present study on $\beta = 0$, but additionally test the robustness for varying values of β ($\beta = 0, 0.25, 0.5, 0.75, 1$).

For a more detailed description of the LRP algorithm, we kindly refer the reader to Bach et al. (2015); Montavon et al. (2018). A PyTorch implementation of the LRP algorithm has been developed for the current work and is available on github[4].

2.3.2. Guided Backpropagation (GB)
In order to emphasize and point out the advantages of LRP as a diagnostic tool, we compared it to a gradient-based method, the guided backpropagation (GB) algorithm (Springenberg et al., 2014). In GB, the absolute values of the gradient of the output with respect to the input nodes is shown as a heatmap, with the additional twist that negative gradients are set to zero at the rectification layers of the network. As was shown by Rieke et al. (2018), "rectifying" the gradients in the backward pass leads to more focused heatmaps.

2.4. Analyzing the Classification Decisions
The CNN model was evaluated on each MR image from the test set and, subsequently, both the LRP as well as the GB algorithm were used to produce a heatmap for each MR image. In the case of LRP, we produced separate heatmaps for each β value. We analyzed the resulting heatmaps (1) group-wise to distill those regions, which are particularly "important" for the AD classification and (2) individually to understand the network decisions per sample and find differences between subjects. For the former, we computed an average AD heatmap (obtained from all AD subjects) and an average HC heatmap (obtained from all HCs), which we then further split into a true positive heatmap (i.e., average heatmap of clinically validated AD patients, who are classified as AD), a false positive heatmap (i.e., average heatmap of HCs classified as AD), a true negative heatmap (i.e., average heatmap of HCs classified as HC) and a false negative heatmap (i.e., average heatmap of clinically validated AD patients classified as HC). For GB, these heatmaps highlight those areas to which the network is on average most susceptible. For LRP, they show the average relevance of each voxel for contributing to the AD score. All LRP heatmaps show the average relevance for the same class (AD), such that they can be compared on the same scale (relevance for AD diagnosis). As the AD scores of HCs

typically range between 0 and 0.5, there will be relevance for AD in HCs, too.

2.5. Atlas-Based Importance Metrics
To quantitatively analyze the heatmaps and the underlying CNN model, we assessed the importance of different brain areas—as defined by the Neuromorphometrics brain atlas (Bakker et al., 2015)—by using the following three metrics for both LRP and GB.

2.5.1. Sum of AD Importance per Area
As a first metric of importance, the resulting heatmap values were simply summed per area. While this can already be taken as a measure of importance, the resulting importance scores are highly correlated to the area size, see **Figure 4**. Therefore, two size-independent metrics for importance were additionally analyzed in more detail: the size-normalized sum, and the average gain (ratio) when comparing to the average HC patient.

2.5.2. Size-Normalized AD Importance Metric
For diagnostic purposes, it can be particularly interesting to identify areas that over their entire volume carry a lot of information, i.e., areas with high *relevance density* or, in GB, *susceptibility density*. Therefore, we divided here the sum of AD importance per area by the size of the area (i.e., number of voxels), which corresponds to the regional mean relevance/susceptibility. While low values over large areas might be due to statistical fluctuations in the data, clusters of relevance (LRP) or susceptibility (GB) in a very confined area could be indicative of the presence of certain biomarkers for AD.

2.5.3. Gain—Ratio of Values With Respect to the Average HC
Lastly, it is important to note that HCs are not "relevance-free" under the LRP algorithm: HCs might exhibit certain structural elements in their brains that are correlated with the AD diagnosis. While the network might still classify them as HC, these structures lead to a class score greater than zero for virtually every subject. Thus, as an additional metric, we will look at the "gain" in relevance (LRP) and susceptibility (GB) per area, i.e., the ratio to the average HC in that area. By doing this, those areas that differ most between the two cases will be attributed the highest importance.

3. RESULTS

In section 3.1, we compare the heatmaps generated by GB and LRP qualitatively with respect to different β values and different sets of data (AD, HC, true positives, false positives etc., see **Figures 2**, **3**). In section 3.2, we quantitatively compare the heatmaps with respect to the different atlas-based importance metrics (see **Figures 4**–**7**). In section 3.3, we present and discuss the LRP heatmaps of two individual patients (see **Figure 8**) and investigate the association between LRP relevance scores and hippocampal volume as one of the neurobiological key markers of AD (see **Figure 9**).
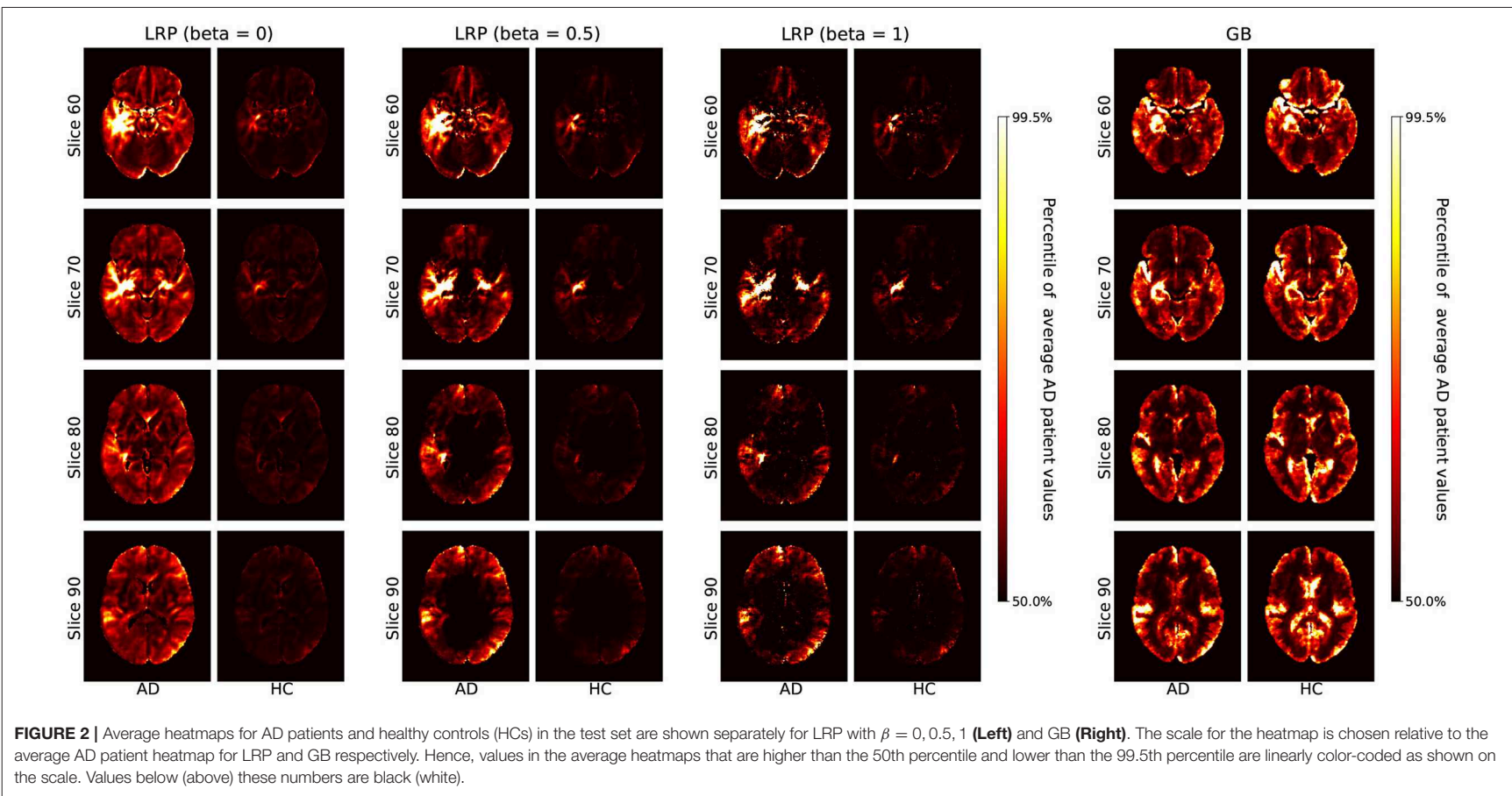
---

[4]https://github.com/moboehle/Pytorch-LRP

FIGURE 2 | Average heatmaps for AD patients and healthy controls (HCs) in the test set are shown separately for LRP with $\beta = 0, 0.5, 1$ **(Left)** and GB **(Right)**. The scale for the heatmap is chosen relative to the average AD patient heatmap for LRP and GB respectively. Hence, 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).
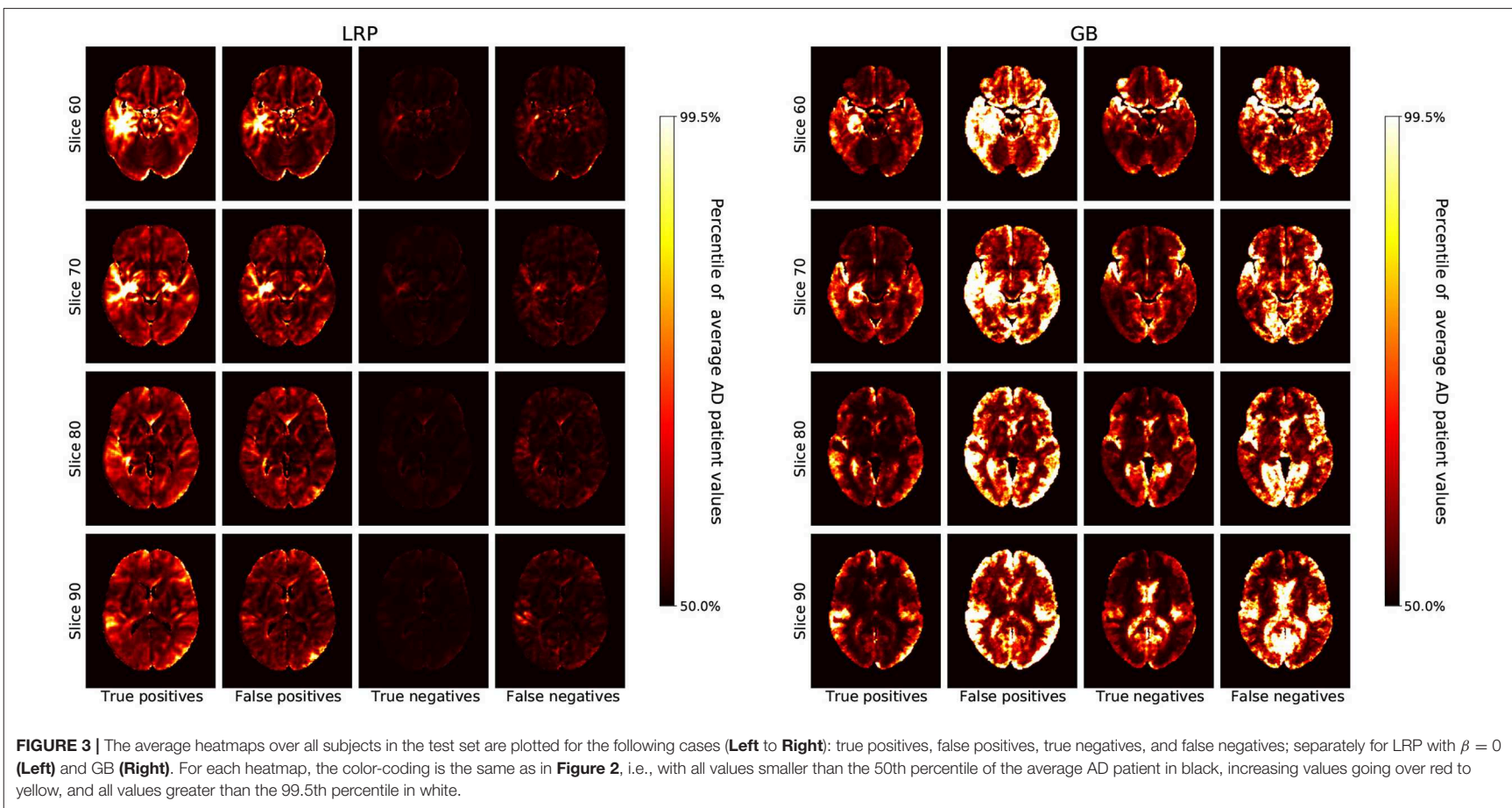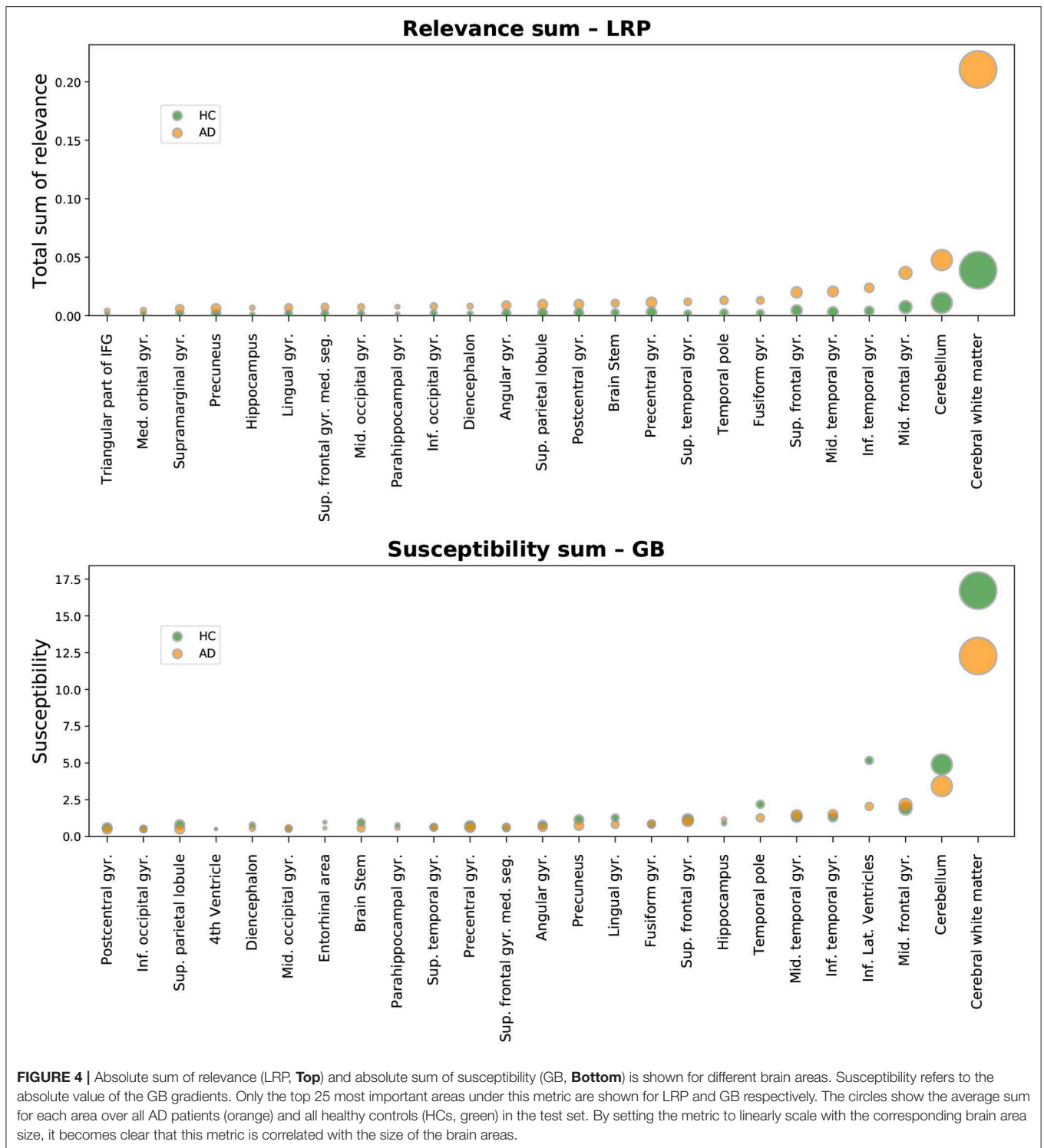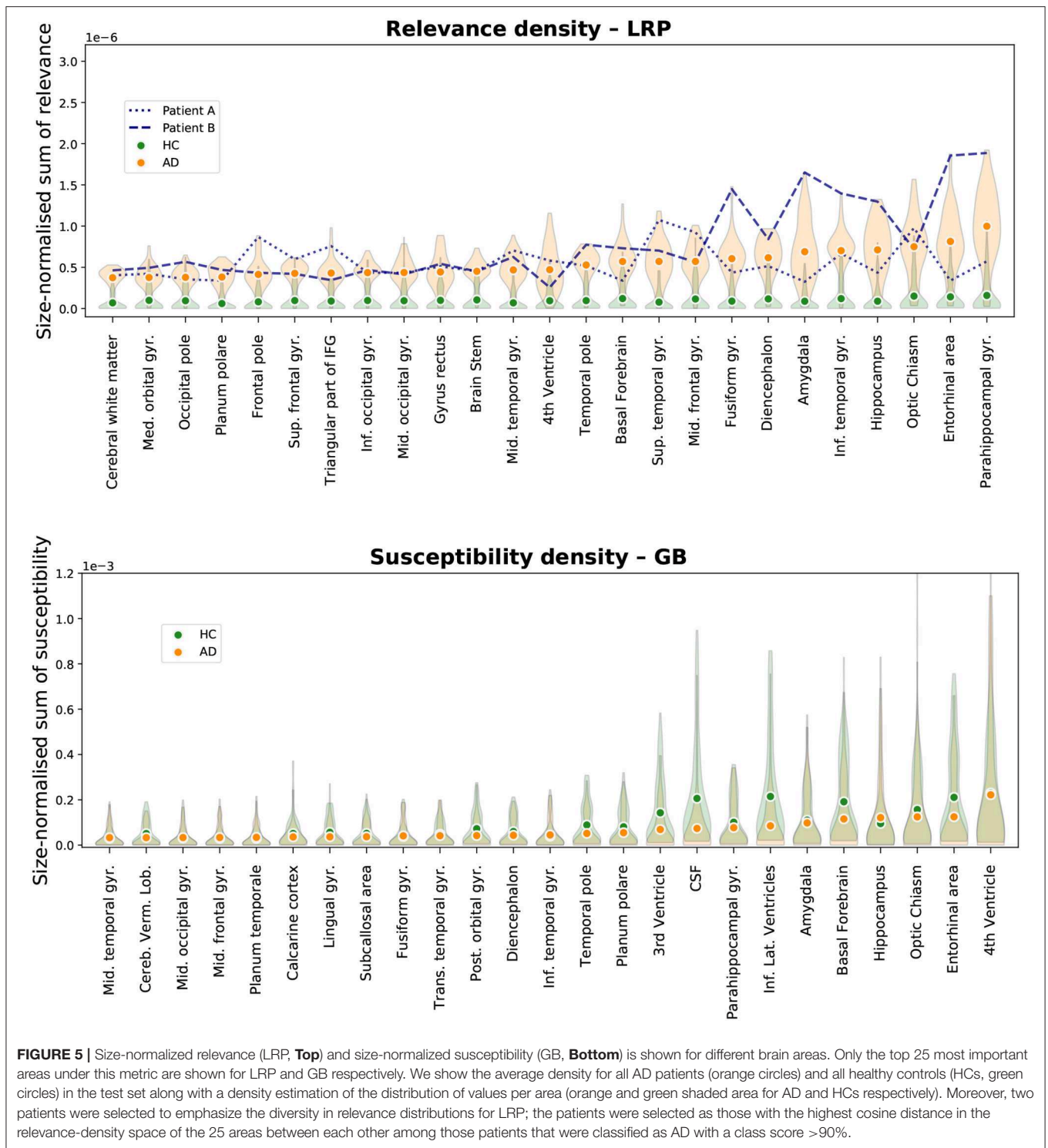
FIGURE 3 | The average heatmaps over all subjects in the test set are plotted for the following cases (**Left** to **Right**): true positives, false positives, true negatives, and false negatives; separately for LRP with $\beta = 0$ **(Left)** and GB **(Right)**. For each heatmap, the color-coding is the same as in **Figure 2**, i.e., with all 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 4 | Absolute sum of relevance (LRP, **Top**) and absolute sum of susceptibility (GB, **Bottom**) is shown for different brain areas. Susceptibility refers to the absolute value of the GB gradients. Only the top 25 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. By setting the metric to linearly scale with the corresponding brain area size, it becomes clear that this metric is correlated with the size of the brain areas.

3.1. Average Heatmap Comparison

In **Figure 2**, we show the average heatmaps for AD patients and HCs, separately for LRP with different β values ($\beta = 0, 0.5, 1$) and GB. The AD pattern between LRP and GB is relatively similar, which is reasonable because all heatmaps are extracted from the same CNN model. However, whereas GB heatmaps are very susceptible for both AD and HCs, LRP heatmaps show much more relevance in AD patients than HCs. This indicates that LRP heatmaps might be more valuable in assessing why a certain person has been classified as AD patient as opposed

FIGURE 5 | Size-normalized relevance (LRP, **Top**) and size-normalized susceptibility (GB, **Bottom**) is shown for different brain areas. Only the top 25 most important areas under this metric are shown for LRP and GB respectively. We show the average density for all AD patients (orange circles) and all healthy controls (HCs, green circles) in the test set along with a density estimation of the distribution of values per area (orange and green shaded area for AD and HCs respectively). Moreover, two patients were selected to emphasize the diversity in relevance distributions for LRP; the patients were selected as those with the highest cosine distance in the relevance-density space of the 25 areas between each other among those patients that were classified as AD with a class score >90%.

to which voxels should be changed to increase the likelihood for AD diagnosis. Concerning the different β values, it is noted that the heatmaps look qualitatively similar, but that sparseness increases with higher β values (which is due to a larger effect of inhibitory contributions, see also Binder et al., 2016a). Since β values close to 0 focus on positive AD contributions and are

thus clinically better interpretable, we focus on $\beta = 0$ in the remaining analyses.

In **Figure 3**, we show the average heatmaps for the distinct classification cases (true positives, false positives etc.), separately for LRP ($\beta = 0$) and GB. In particular, the false positives lead to an interesting insight: For LRP, the false positives
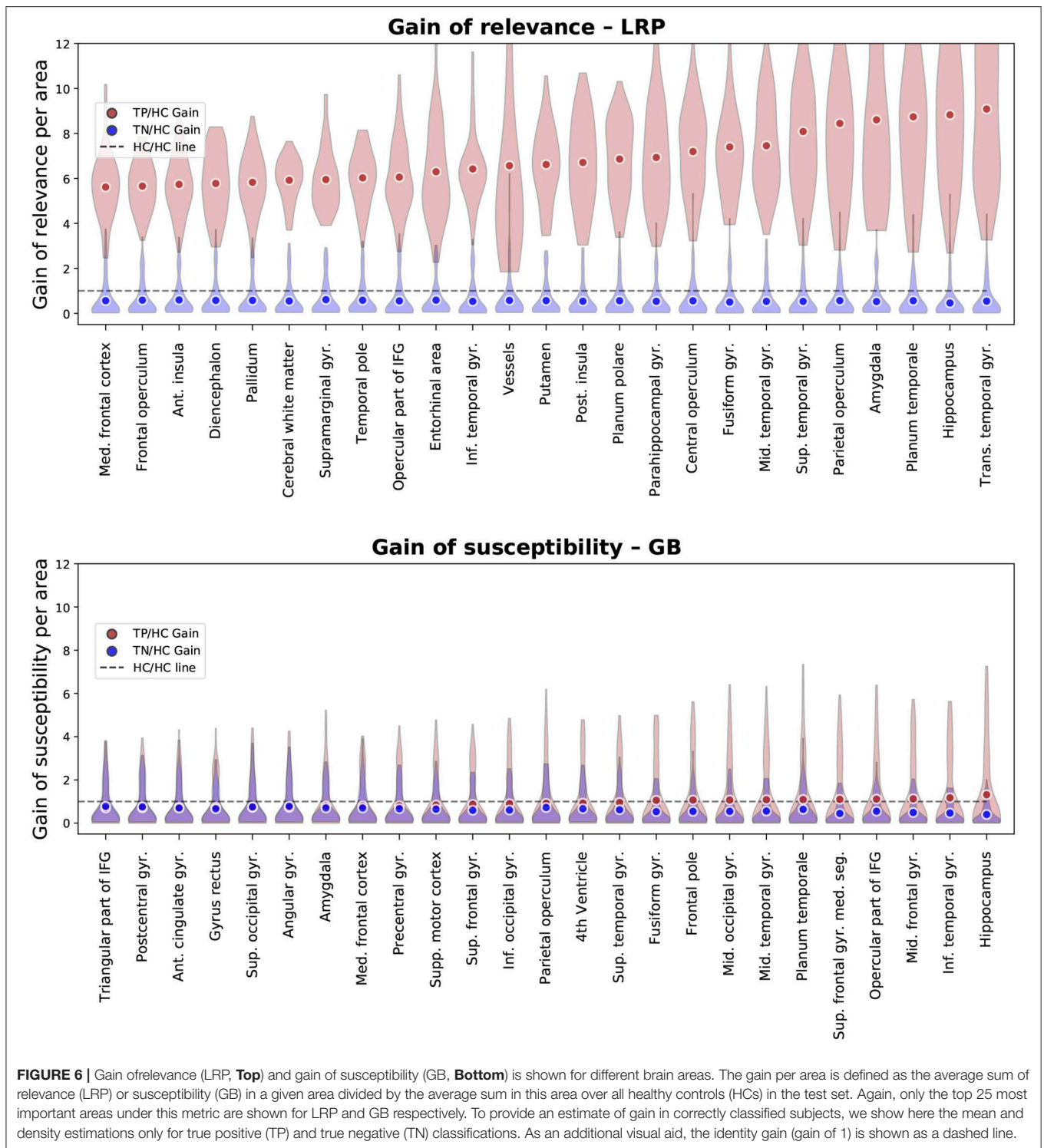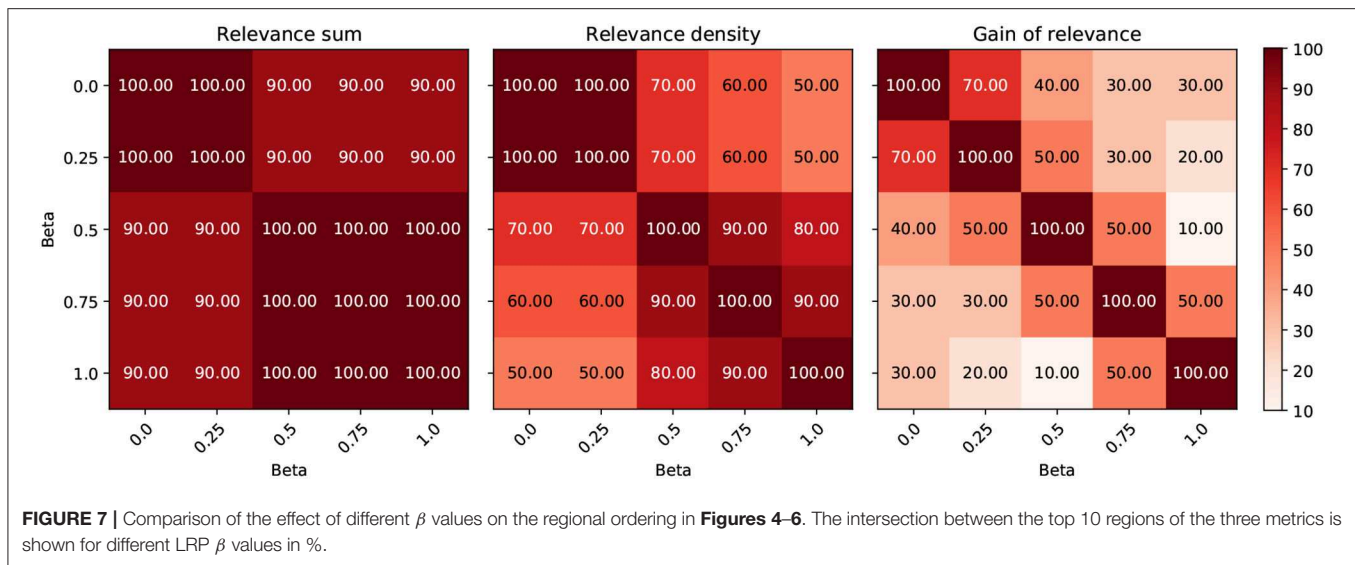
FIGURE 6 | Gain of relevance (LRP, **Top**) and gain of susceptibility (GB, **Bottom**) is shown for different brain areas. The gain per area is defined as the average sum of relevance (LRP) or susceptibility (GB) in a given area divided by the average sum in this area over all healthy controls (HCs) in the test set. Again, only the top 25 most important areas under this metric are shown for LRP and GB respectively. To provide an estimate of gain in correctly classified subjects, we show here the mean and density estimations only for true positive (TP) and true negative (TN) classifications. As an additional visual aid, the identity gain (gain of 1) is shown as a dashed line.

exhibit less relevance than the true positives, but generally in similar areas. This could indicate that in these patients structures that are correlated with AD were found, albeit that overall the positive contribution was less compelling than for true AD patients. For GB, on the other hand, the false classifications (mostly false positives, but also false negatives) seem to exhibit the highest gradient values of all cases. This exemplifies well what GB truly measures: in the case of false positives (and negatives), the network might be "unsure" and more easily influenced to change its decision; the outcome is unstable. The highlighted areas that could change the outcome are very broadly distributed and need

FIGURE 7 | Comparison of the effect of different β values on the regional ordering in **Figures 4–6**. The intersection between the top 10 regions of the three metrics is shown for different LRP β values in %.

not necessarily represent areas with positive contributions for AD.

3.2. Atlas-Based Importance Metrics

In **Figure 4**, we show the sum of AD importance per area, separately for LRP ($\beta = 0$) and GB. Although this metric seems to be dominated by the size of the respective brain area, one important qualitative difference between LRP and GB is visible: in the LRP results, the mean importance values per area are consistently much higher for AD patients than for HCs. For GB, this clear split is not present; moreover, the average sum of gradients in several brain regions, including the cerebral white matter and cerebellum, is even higher for HC than for AD. This exemplifies well that the heatmaps for GB cannot directly be interpreted as showing the relevance for AD classification, but instead show the sensitivity of the outcome to certain areas, which does not have to be AD or HC specific. As the absolute sum of importance correlates with the size of the respective brain area, the following metrics, in which we controlled for the brain area size, might be better interpretable.

In **Figure 5**, the total sum of importance is normalized by the size of the respective brain area. Here, the aforementioned difference in the distributions between HCs and AD patients becomes even more apparent: while the distributions are very heavily overlapping for GB, this is not the case for LRP. Notably, the variance in the AD distributions is much higher in the AD case than in the HC case. This could indicate that the network has learned to differentiate between subtypes of AD and bases its decision on different structural elements for different patients; the existence of different subtypes of AD has been investigated in recent work, see for example (Ferreira et al., 2017; Park et al., 2017). In contrast, for HCs the relevance density is consistently very low. As an example of the diversity in importance assessments according to this metric, we added the "individual fingerprints" of two AD patients to **Figure 5**; for these
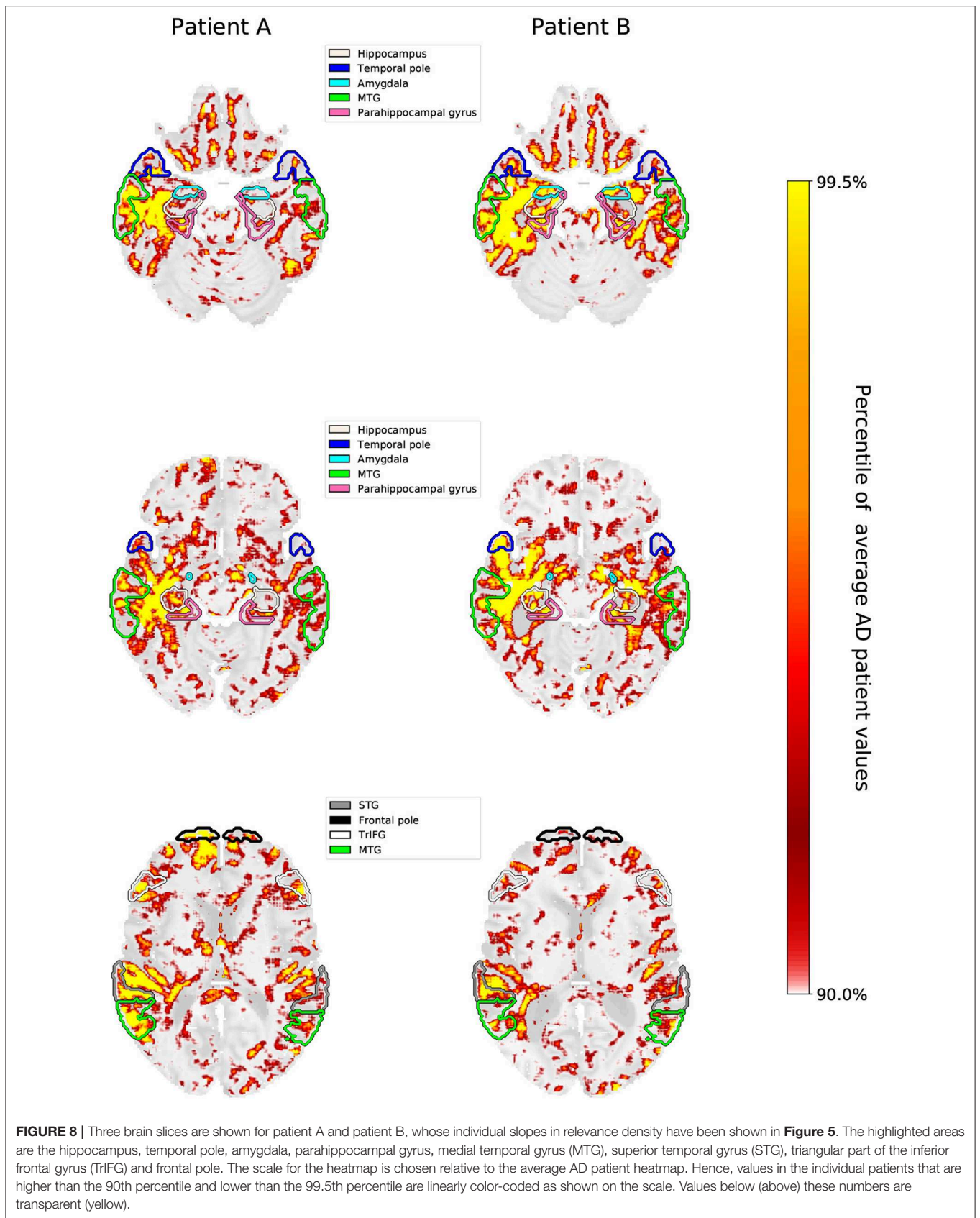
patients the individual heatmaps will be compared in section 3.3 and **Figure 8**.

In **Figure 6**, the results for the gain metric for different cases—true positives and true negatives—are visualized. This metric allows for plotting the LRP and the GB results on the same scale and emphasizes once again the stronger distinction between AD patients and HCs under the LRP algorithm. Most gain for LRP has been found in areas of temporal lobe including transversal temporal gyrus, hippocampus, planum temporale and amygdala.

In **Figure 7**, we compare the regional overlap of the top 10 regions between the β values 0, 0.25, 0.5, 0.75, 1, separately for the three importance metrics. It is shown that (1) the regional overlap is strongest for relevance sum followed by relevance density and relatively unstable for gain of relevance especially for large and more distant β values and (2) the regional overlap is–as expected–stronger for neighboring β values. The instability of the gain metric for higher β values is probably due to the associated sparsity leading to very low relevance scores for HCs (which might—in some cases—inflate the gain metric).

3.3. Individual Heatmaps—Fingerprinting and Neurobiological Relevance

Since the LRP heatmaps take into account the individual filter activations and therefore highlight positive contributions to the class score of AD, they might serve as "individual fingerprints" in a diagnostic tool. In **Figure 8**, we show several slices of the relevance heatmaps for two patients in order to highlight the diversity in those heatmaps. The two patients were selected from the test set as those with the highest cosine distance in the relevance-density space between each other among those patients that were classified as AD with a class score > 90% (their individual trajectories of region-wise relevance are shown in **Figure 5**). It can be seen that the areas, which mainly contributed to the network decision, are rather different for the two patients. For one patient (patient B), the class score of the network is heavily influenced by areas of the temporal lobe, such as
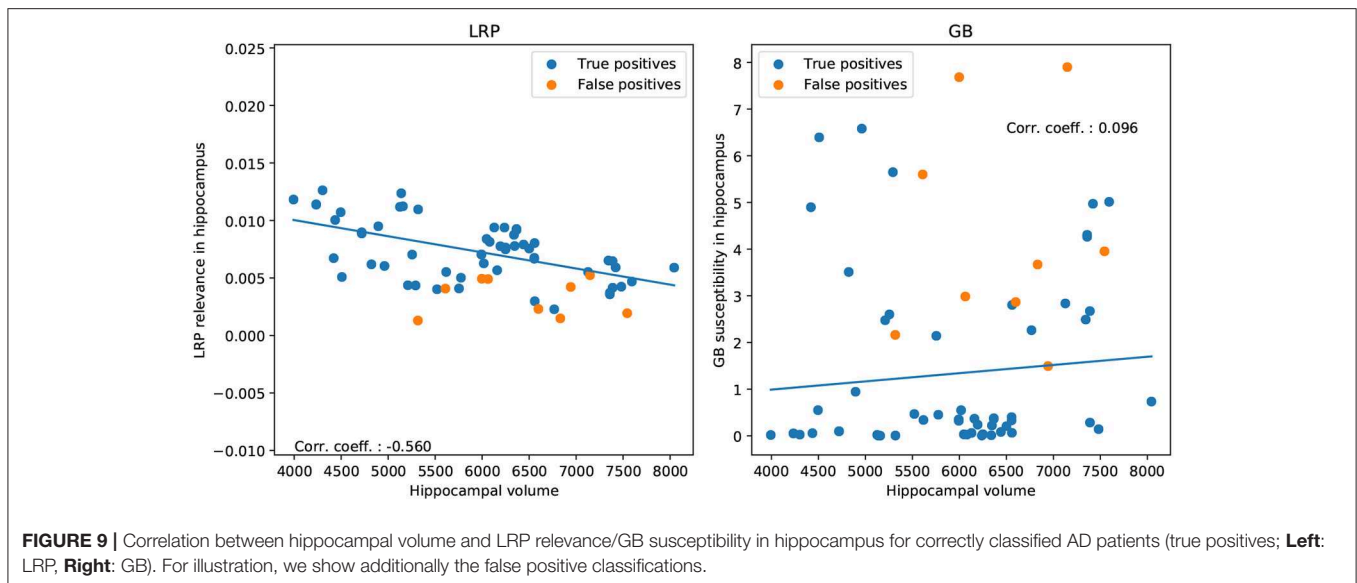
FIGURE 8 | Three brain slices are shown for patient A and patient B, whose individual slopes in relevance density have been shown in **Figure 5**. The highlighted areas are the hippocampus, temporal pole, amygdala, parahippocampal gyrus, medial temporal gyrus (MTG), superior temporal gyrus (STG), triangular part of the inferior frontal gyrus (TrIFG) and frontal pole. 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 9 | Correlation between hippocampal volume and LRP relevance/GB susceptibility in hippocampus for correctly classified AD patients (true positives; **Left**: LRP, **Right**: GB). For illustration, we show additionally the false positive classifications.

parahippocampal gyrus, entorhinal area, hippocampus, inferior temporal gyrus and amygdala, while for the second patient (patient A), frontal areas, including triangular part of inferior frontal gyrus, superior frontal gyrus and frontal pole, in addition to superior temporal gyrus seem to be most informative.

To investigate whether higher importance scores correspond to stronger anatomical deviations (e.g., atrophy) in correctly classified AD patients (true positives), we performed a correlation analysis between hippocampal volume and LRP relevance/GB susceptibility scores (see **Figure 9**). We show that the LRP relevance score ($\beta = 0$) in the hippocampus is significantly (negatively) correlated with hippocampal volume ($-0.560, p < 10^{-3}$, permutation test), whereas the GB score is not ($0.096, p = 0.77$). To rule out that false positives are outliers in terms of association between hippocampal volume and LRP relevance, we included them in **Figure 9**. Interestingly, for larger β values the correlation tends to decrease ($-0.560, -0.562, -0.525, -0.457, -0.361$ for $\beta = 0, 0.25, 0.5, 0.75, 1$ respectively) supporting our notion of a higher neurobiological relevance in case of β values close to 0.

4. DISCUSSION

In this study, we introduced LRP as a powerful method for explaining individual CNN decisions in AD classification. After training a CNN to separate AD patients and HCs based on structural MRI data, individual heatmaps—indicating the importance for each voxel for the respective classification decision—were produced for the test subjects. We analyzed the heatmaps with respect to different classification subgroups (AD patients, HCs, true positives, false positives etc.) and different β values. The relevance of brain regions contained in the Neuromorphometrics atlas was evaluated using three different importance metrics, namely the sum of importance per area, the size-normalized AD importance, and the gain as ratio between

AD and HC importance. We demonstrated that LRP-derived heatmaps—in contrast to GB—provide (1) high specificity for individuals and (2) little relevance for AD in HCs. Additionally, areas that exhibit a lot of relevance correlate well with what is known from literature. Importantly, these LRP heatmaps were produced without the need for expert annotations on the presence or absence of biomarkers throughout the learning process. This combination of a simple classification task (AD vs. HC) and in-depth network analysis by LRP might be a promising tool for diagnostics. Additionally, it could allow for discovering new and unknown biomarkers for a variety of diseases and might help distinguishing subtypes of AD by analyzing the diversity in "relevance hot-spots" across all AD patients. Furthermore, the size-corrected metrics ("relevance density" and "relevance gain") seem to correlate well with what is known from AD research, indicating that the most discriminating features for classifying an input image as AD can be found in the temporal lobe. We therefore think that a well-trained neural network, analyzed by means of the LRP algorithm, can become a useful tool for practitioners and increase the trust in computer-aided diagnoses, as an interpretable explanation of the decision can be produced.

4.1. Regional Specificity of LRP

We quantitatively evaluated the heatmaps, obtained by either GB or LRP, toward different brain areas according to the Neuromorphometrics atlas (Bakker et al., 2015) by summarizing the importance (AD relevance in case of LRP, susceptibility in case of GB) for each brain area separately. Both types of heatmaps mostly identified regions known to be important in disease progression of AD, such as structures in the medial temporal lobe including hippocampus, amygdala, parahippocampal gyrus, and entorhinal cortex (Du et al., 2001; Desikan et al., 2009; Frisoni et al., 2010; Velayudhan et al., 2013; Weiner et al., 2013; Klein-Koerkamp et al., 2014; Long et al., 2017) as well as frontal and parietal areas (Casanova et al., 2011; Quiroz et al., 2013; Kilimann et al., 2017; Park et al., 2017; Liu et al., 2018). For

all these regions morphometric changes including global and local atrophy (e.g., smaller volumes of hippocampus or amygdala, reduced cortical thickness or gray matter density) or deviations in shape have been shown and related to disease progression and cognitive decline (Desikan et al., 2009; Frisoni et al., 2010; Weiner et al., 2013; Hidalgo-Muñoz et al., 2014; Long et al., 2017; Ledig et al., 2018). These changes seem to be utilized by our CNN framework for making individual predictions and are highlighted in the heatmaps of both LRP and GB. However, the contrast in importance scores between AD patients and HCs is much higher for LRP than GB (in GB, the average heatmaps for AD patients and HCs are quite similar). This supports the notion that LRP heatmaps reflect AD-specific relevance, whereas GB emphasizes areas which the network more generally is sensitive to. Regarding other structures found to be important in our network, it might be interesting to see if also other neural networks find relevance in these areas and if predictions about finding significant structural changes in these areas might be possible at some point. In this respect, the decisions of such networks can be treated as a "second opinion" and a reciprocal learning process with medical experts might be initiated.

4.2. Fingerprinting and Neurobiological Relevance

In addition to heatmap differences between AD patients and HCs, we noticed a high variability between the heatmaps of individual AD patients for the LRP method. This variability was not only reflected in a high variance of important scores within regions, but also in individual trajectories ("fingerprints"), which we exemplary depicted for two AD patients, see **Figure 8**. For future work, it might be very interesting to see if these fingerprints reflect different disease stages of AD (Braak and Braak, 1991; Casanova et al., 2011) or allow for identifying subtypes of AD, in which brain areas are affected differently (Murray et al., 2011; Noh et al., 2014; Scheltens et al., 2016; Zhang et al., 2016; Ferreira et al., 2017; Park et al., 2017). Zhang et al. (2016), for example, identified a temporal, a subcortical and a cortical atrophy factor associated with impairment in different cognitive domains. Another important question is whether the relevance found by the LRP method reflect some true evidence in the sense of biomarkers. By showing that the hippocampal volume is significantly correlated to the LRP relevance scores (but not to the GB susceptibility scores), we argue that LRP—at least partially—succeeded here in breaking down the relevance to the level of voxels in a meaningful way. Interestingly, we found higher correlations for lower β values speaking for a higher neurobiological relevance of β values close to 0. Further studies are needed to more carefully relate LRP relevance measures to other clinical markers of AD including biomarkers and neuropsychological test scores, also in dependency of different CNN models and parameter settings. Moreover, our metrics should be evaluated in patients with mild cognitive impairment (MCI).

4.3. Related Work

Visualization of deep neural networks is a fairly new research area and different attempts have been made to provide intuitive explanations for neural network decisions. However, there is not yet a state-of-the art visualization method as saliency maps for example have been shown to be misleading (Adebayo et al., 2018). In Alzheimer's research, there are only a couple of studies that looked into different visualization methods based on MRI and/or PET data. Most of these studies either visualized filters and activations of the first or last layer (Sarraf and Tofighi, 2016; Lu et al., 2018; Ding et al., 2019) or used the occlusion method to exclude some parts (e.g., with a black patch) of the input image and recalculate the classifier output (Korolev et al., 2017; Esmaeilzadeh et al., 2018; Liu et al., 2018). Based on visual impression, they found that the networks focus primarily on areas known to be involved in AD, such as hippocampus, amygdala or ventricles, but occasionally also other areas such as thalamus or parietal lobe appear. Importantly, in contrast to our study, they did not quantitatively analyze the data, e.g., with respect to brain areas contained in an atlas or underlying neurobiological markers. Additionally, they did not compare different visualization methods or looked for inter-individual differences. One study, however, used gradient-weighted classification activation mapping (grad-CAM) and compared it to sensitivity analysis for AD classification (Yang et al., 2018). They demonstrate that these different visualization methods capture different aspects of the data and show high variability depending e.g., on the resolution of the convolutional layers. In Rieke et al. (2018), gradient-based and occlusion methods (standard patch occlusion and brain area occlusion) were qualitatively and quantitatively compared for AD classification. High regional overlaps between the methods, mostly inferior and middle temporal gyrus, were found but for gradient-based methods the importance was more widely distributed. Regarding the LRP method, we are only aware of one application in the neuroimaging field: Thomas et al. (2018) introduce interpretable recurrent networks for decoding cognitive states based on functional MRI data and demonstrate that the LRP method is capable of identifying relevant brain areas for the different tasks and different levels of data granularity.

4.4. Limitations

Although LRP heatmaps seem to be a promising tool for visualizing neural network decisions, we would like to point out several limitations of LRP and other heatmap methods in the context of this study.

First, heatmap methods are limited by the lack of a ground truth. Most commonly, heatmaps are qualitatively evaluated based on visual assessment, but there are also studies proposing sanity checks (Adebayo et al., 2018) or more objective quality measures such as region perturbation (Samek et al., 2015). In Lipton (2018), the interpretability of models has been generally investigated and questioned. In medical research, heatmaps can be qualitatively evaluated based on prior knowledge (e.g., hippocampus is known to be strongly affected in AD, therefore it seems reasonable to find relevance there). Given that in the specific case of heatmaps for MR images the input space is highly structured, we proposed here additional ways for assessing the quality of explanations by using a brain atlas. Future

studies might assess the neurobiological validity by removing presumably important brain areas and re-training the classifier.

Second, it is largely acknowledged that heatmaps are quite sensitive to the specific algorithms (and its parameters, e.g., the β value in case of LRP) used to produce them. However, regarding the β values in LRP, we have shown that the heatmaps are relatively robust toward this parameter, only sparsity increases as a function of β. Additionally, we demonstrated that the regional ordering is relatively stable for relevance sum and density, but unstable for the gain metric—especially in the case of large and more distant β values.

Third, heatmaps just highlight voxels that contributed to a certain classifier decision, but do not allow making a statement about the underlying reasons (e.g., atrophy or shape differences) or potential interactions between voxels or brain areas. For example, it is difficult to disentangle interactions between different regions (certain patterns in the hippocampus might only be considered as positive evidence if structure Y is found in area Z) nor do we know whether the network developed specific filters for atrophy or the shapes of different structures. Although we found in this study a significant correlation between hippocampal volume and LRP relevance measures, we can not make any claim about causal relationships here. Future studies are necessary to more systematically investigate the relationship between manifested neurobiological markers and LRP explanations.

Fourth, heatmaps strongly depend on the type and quality of the classifier, whose decisions are sought to be explained. Therefore, each heatmap should be read as an indication of where the specific network model sees evidence. For badly trained networks, this does not have to correlate at all with the presence of actual biomarkers. Nevertheless, the better the classifier, the more likely it becomes that the classifier uses meaningful patterns as a basis for its decision and that the heatmaps correlate with "true" evidence for AD. However, heatmaps are also useful in cases, where classification performance is low or sample size is rather small, e.g., for better understanding if the classifier picks up relevant or irrelevant features (e.g., noise or imaging artifacts) and if there are any biases present in the data set (Lapuschkin et al., 2016; Montavon et al., 2018). It would be very interesting to investigate how the heatmaps change for different networks, as those which yield stronger classification results should also base their decisions on *better* "evidence".

And finally, it should be stressed that when we refer to brain areas throughout this work, we refer to the location that the areas are assigned in the brain atlas and not to the individual anatomical structures of any patient. Due to inter-individual differences, the match between the atlas and the individual patient's anatomical realities will not be perfect; this is most likely further aggravated by the presence of atrophy in AD patients.

5. CONCLUSION

In conclusion, we introduced the LRP method for explaining individual CNN decisions in MRI-based AD diagnosis. In contrast to GB, LRP heatmaps can be interpreted as providing individual AD relevance ("What speaks for AD in this particular subject?") as opposed to a general susceptibility for small variations in the input data. Additionally, we provided a framework and specific metrics (i.e., "relevance density" and "relevance gain") to quantitatively compare heatmaps between different groups, brain areas or methods. We demonstrated that these metrics correlate well with clinical findings in AD, but also vary strongly between AD patients. By this, the LRP method might be very useful in a clinical setting for a case-by-case evaluation. However, we would like to point out that (1) our metrics should be evaluated in different network architectures and (2) other (individual) brain atlases might be used for the evaluation of regions. Future studies should evaluate the LRP method on patients with mild-cognitive impairment (MCI) and relate findings to known biomarkers in AD. We are convinced that our framework might also be very useful for other disease classification studies in helping to understand individual network decisions.

DATA AVAILABILITY

The ADNI data set is for researchers publicly available at http://adni.loni.usc.edu/. The code is available at https://github.com/moboehle/Pytorch-LRP.

AUTHOR CONTRIBUTIONS

MB, FE, MW, and KR designed the study. MB and FE engineered the software and analyzed the data. MB, FE, and KR wrote the paper.

FUNDING

ACKNOWLEDGMENTS

REFERENCES

Abdel-Hamid, O., Mohamed, A.-R., Jiang, H., Deng, L., Penn, G., and Yu, D. (2014). Convolutional neural networks for speech recognition. *IEEE/ACM Trans. Audio Speech. Language Proc.* 22, 1533–1545. doi: 10.1109/TASLP.2014.2339736

Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., and Kim, B. (2018). "Sanity checks for saliency maps," in *Advances in Neural Information Processing Systems*, eds S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (Montréal, QC: Curran Associates, Inc.), 9505–9515.

Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.-R., and Samek, W. (2015). On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLoS ONE* 10:e0130140. doi: 10.1371/journal.pone.0130140

Bakker, R., Tiesinga, P., and Kötter, R. (2015). The scalable brain atlas: instant web-based access to public brain atlases and related content. *Neuroinformatics* 13, 353–366. doi: 10.1007/s12021-014-9258-x

Binder, A., Bach, S., Montavon, G., Müller, K.-R., and Samek, W. (2016a). "Layer-wise relevance propagation for deep neural network architectures," in *Information Science and Applications (ICISA) 2016*, eds K. J. Kim and N. Joukov (Singapore: Springer Singapore), 913–922.

Binder, A., Montavon, G., Lapuschkin, S., Müller, K.-R., and Samek, W. (2016b). "Layer-wise relevance propagation for neural networks with local renormalization layers," in *International Conference on Artificial Neural Networks*, eds A. E. P. Villa, P. Masulli, and A. J. P. Rivero (Barcelona: Springer), 63–71.

Bondi, M. W., Edmonds, E. C., and Salmon, D. P. (2017). Alzheimer's disease: past, present, and future. *J. Int. Neuropsychol. Soc.* 23, 818–831. doi: 10.1017/S135561771700100X

Braak, H., and Braak, E. (1991). Neuropathological stageing of Alzheimer-related changes. *Acta Neuropathol.* 82, 239–259.

Casanova, R., Whitlow, C. T., Wagner, B., Williamson, J., Shumaker, S. A., Maldjian, J. A., et al. (2011). High dimensional classification of structural MRI Alzheimer's Disease data based on large scale regularization. *Front. Neuroinformat.* 5:22. doi: 10.3389/fninf.2011.00022

Castelvecchi, D. (2016). Can we open the black box of AI? *Nature* 538, 20–23. doi: 10.1038/538020a

Desikan, R. S., Cabral, H. J., Hess, C. P., Dillon, W. P., Glastonbury, C. M., Weiner, M. W., et al. (2009). Automated MRI measures identify individuals with mild cognitive impairment and Alzheimer's disease. *Brain* 132, 2048–2057. doi: 10.1093/brain/awp123

Ding, Y., Sohn, J. H., Kawczynski, M. G., Trivedi, H., Harnish, R., Jenkins, N. W., et al. (2019). A deep learning model to predict a diagnosis of Alzheimer disease by using 18 F-FDG PET of the brain. *Radiology* 290, 456–464. doi: 10.1148/radiol.2018180958

Du, A. T., Schuff, N., Amend, D., Laakso, M. P., Hsu, Y. Y., Jagust, W. J., et al. (2001). Magnetic resonance imaging of the entorhinal cortex and hippocampus in mild cognitive impairment and Alzheimer's disease. *J. Neurol. Neurosurg. Psychiat.* 71, 441–447. doi: 10.1136/jnnp.71.4.441

Esmaeilzadeh, S., Belivanis, D. I., Pohl, K. M., and Adeli, E. (2018). "End to-end Alzheimer's disease diagnosis and biomarker identification," in *International Workshop on Machine Learning in Medical Imaging*, eds Y. ShiHeung-Il and S. Liu (Granada: Springer), 337–345.

Ferreira, D., Verhagen, C., Hernández-Cabrera, J. A., Cavallin, L., Guo, C.-J., Ekman, U., et al. (2017). Distinct subtypes of Alzheimer's disease based on patterns of brain atrophy: longitudinal trajectories and clinical applications. *Sci. Rep.* 7:46263. doi: 10.1038/srep46263

Frisoni, G. B., Fox, N. C., Jack C. R. Jr., Scheltens, P., and Thompson, P. M. (2010). The clinical use of structural MRI in Alzheimer disease. *Nat. Rev. Neurol.* 6, 67–77. doi: 10.1038/nrneurol.2009.215

Gupta, A., Ayhan, M., and Maida, A. (2013). "Natural image bases to represent neuroimaging data," in *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, Vol 28, eds S. Dasgupta and D. Mcallester (Atlanta, GA: JMLR Workshop and Conference Proceedings), 987–994.

Hidalgo-Muñoz, A. R., Ramírez, J., Górriz, J. M., and Padilla, P. (2014). Regions of interest computed by SVM wrapped method for Alzheimer's disease examination from segmented MRI. *Front. Aging Neurosci.* 6:20. doi: 10.3389/fnagi.2014.00020

Kilimann, I., Hausner, L., Fellgiebel, A., Filippi, M., Würdemann, T. J., Heinsen, H., et al. (2017). Parallel atrophy of cortex and basal forebrain cholinergic system in mild cognitive impairment. *Cereb. Cortex* 27, 1841–1848. doi: 10.1093/cercor/bhw019

Kingma, D. P., and Ba, J. (2015). *Adam: A Method for Stochastic Optimization.* San Diego, CA: ICLR.

Klein-Koerkamp, Y., Heckemann, R., Ramdeen, K., Moreaud, O., Keignart, S., Krainik, A., et al. (2014). Amygdalar atrophy in early Alzheimer's disease. *Curr. Alzheimer Res.* 11, 239–252. doi: 10.2174/1567205011666140131123653

Klöppel, S., Stonnington, C. M., Chu, C., Draganski, B., Scahill, R. I., Rohrer, J. D., et al. (2008). Automatic classification of MR scans in Alzheimer's disease. *Brain* 131, 681–689. doi: 10.1093/brain/awm319

Korolev, S., Safiullin, A., Belyaev, M., and Dodonova, Y. (2017). "Residual and plain convolutional neural networks for 3d brain mri classification," in *2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017)* (Melbourne, VIC: IEEE), 835–838.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). "ImageNet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems 25*, eds F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger (Lake Tahoe, CA: Curran Associates, Inc), 1097–1105.

Lapuschkin, S., Binder, A., Montavon, G., Muller, K.-R., and Samek, W. (2016). "Analyzing classifiers: fisher vectors and deep neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Las Vegas, CA), 2912–2920.

LeCun, Y., and Bengio, Y. (1995). "Convolutional networks for images, speech, and time series," in *The Handbook of Brain Theory and Neural Networks*, ed M. A. Arbib (MIT Press), 3361.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature* 521:436. doi: 10.1038/nature14539

Ledig, C., Schuh, A., Guerrero, R., Heckemann, R. A., and Rueckert, D. (2018). Structural brain imaging in Alzheimer's disease and mild cognitive impairment: biomarker analysis and shared morphometry database. *Sci. Rep.* 8:11258. doi: 10.1038/s41598-018-29295-9

Lipton, Z. C. (2018). "The mythos of model interpretability," in *Proceedings of the 2016 ICML Workshop on Human Interpretability in Machine Learning*, eds B. Kim, D. M. Malioutov, and K. R. Varshney (New York, NY), 96–100.

Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., et al. (2017). A survey on deep learning in medical image analysis. *Med. Image Analys.* 42, 60–88. doi: 10.1016/j.media.2017.07.005

Liu, M., Cheng, D., Wang, K., and Wang, Y. (2018). Multi-modality cascaded convolutional neural networks for Alzheimer's disease diagnosis. *Neuroinformatics* 16, 295–308. doi: 10.1007/s12021-018-9370-4

Long, J., Shelhamer, E., and Darrell, T. (2015). "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition* (Boston, MA), 3431–3440.

Long, X., Chen, L., Jiang, C., Zhang, L., Initiative, A. D. N., et al. (2017). Prediction and classification of Alzheimer disease based on quantification of MRI deformation. *PLoS ONE* 12:e0173372. doi: 10.1371/journal.pone.0173372

Lu, D., Popuri, K., Ding, G. W., Balachandar, R., Beg, M. F., and Alzheimer's Disease Neuroimaging Initiative. (2018). Multimodal and Multiscale Deep Neural Networks for the Early Diagnosis of Alzheimer's Disease using structural MR and FDG-PET images. *Sci. Rep.* 8:5697. doi: 10.1038/s41598-018-22871-z

Montavon, G., Samek, W., and Müller, K. R. (2018). Methods for interpreting and understanding deep neural networks. *Digital Signal Proc.* 73, 1–15. doi: 10.1016/j.dsp.2017.10.011

Morris, J. C. (1993). The Clinical Dementia Rating (CDR). *Neurology* 43, 2412–2412–a.

Murray, M. E., Graff-Radford, N. R., Ross, O. A., Petersen, R. C., Duara, R., and Dickson, D. W. (2011). Neuropathologically defined subtypes of Alzheimer's disease with distinct clinical characteristics: a retrospective study. *Lancet Neurol.* 10, 785–796. doi: 10.1016/S1474-4422(11)70156-9

Noh, Y., Jeon, S., Lee, J. M., Seo, S. W., Kim, G. H., Cho, H., Ye, B. S., et al. (2014). Anatomical heterogeneity of Alzheimer disease based on cortical thickness on MRIs. *Neurology* 83, 1936–1944. doi: 10.1212/WNL.0000000000001003

Orrù, G., Pettersson-Yeo, W., Marquand, A. F., Sartori, G., and Mechelli, A. (2012). Using support vector machine to identify imaging biomarkers of neurological and psychiatric disease: a critical review. *Neurosci. Biobehav. Rev.* 36, 1140–1152. doi: 10.1016/j.neubiorev.2012.01.004

Park, J.-Y., Na, H. K., Kim, S., Kim, H., Kim, H. J., Seo, S. W., et al. (2017). Robust identification of Alzheimer's disease subtypes based on cortical atrophy patterns. *Sci. Rep.* 7:43270. doi: 10.1038/srep43270

Payan, A., and Montana, G. (2015). Predicting Alzheimer's disease: a neuroimaging study with 3D convolutional neural networks. *CoRR* abs/1502.0.

Quiroz, Y. T., Stern, C. E., Reiman, E. M., Brickhouse, M., Ruiz, A., Sperling, R. A., et al. (2013). Cortical atrophy in presymptomatic Alzheimer's disease presenilin 1 mutation carriers. *J. Neurol. Neurosurg. Psychiat.* 84, 556–561. doi: 10.1136/jnnp-2012-303299

Rathore, S., Habes, M., Iftikhar, M. A., Shacklett, A., and Davatzikos, C. (2017). A review on neuroimaging-based classification studies and associated feature extraction methods for Alzheimer's disease and its prodromal stages. *NeuroImage* 155, 530–548. doi: 10.1016/j.neuroimage.2017.03.057

Rieke, J., Eitel, F., Weygandt, M., Haynes, J.-D., and Ritter, K. (2018). "Visualizing convolutional networks for MRI-based diagnosis of Alzheimer's disease. in *Understanding and Interpreting Machine Learning in Medical Image Computing Applications* (Granada: Springer), 24–31.

Ritter, K., Lange, C., Weygandt, M., Mäurer, A., Roberts, A., Estrella, M., et al. (2016). Combination of structural MRI and FDG-PET of the brain improves diagnostic accuracy in newly manifested cognitive impairment in geriatric inpatients. *J. Alzheimer's Dis.* 54, 1319–1331. doi: 10.3233/JAD-160380

Ritter, K., Schumacher, J., Weygandt, M., Buchert, R., Allefeld, C., and Haynes, J.-D. (2015). Multimodal prediction of conversion to Alzheimer's disease based on incomplete biomarkers. *Alzheimer's Dement. Diagnos. Assess. Dis. Monitor.* 1, 206–215. doi: 10.1016/j.dadm.2015.01.006

Samek, W., Binder, A., Montavon, G., Bach, S., and Müller, K.-R. (2015). Evaluating the visualization of what a deep neural network has learned. *arXiv* arXiv:1509.06321.

Sarraf, S., and Tofighi, G. (2016). DeepAD: Alzheimer's disease classification via deep convolutional neural networks using MRI and fMRI. *bioRxiv*.

Scheltens, N. M., Galindo-Garre, F., Pijnenburg, Y. A., van der Vlies, A. E., Smits, L. L., Koene, T., et al. (2016). The identification of cognitive subtypes in Alzheimer's disease dementia using latent class analysis. *J. Neurol. Neurosurg. Psychiatry* 87, 235–243. doi: 10.1136/jnnp-2014-309582

Simonyan, K., Vedaldi, A., and Zisserman, A. (2013). Deep inside convolutional networks: visualising image classification models and saliency maps. *arXiv* arXiv:1312.6034.

Springenberg, J. T., Dosovitskiy, A., Brox, T., and Riedmiller, M. (2014). Striving for simplicity: the all convolutional net. *arXiv* arXiv:1412.6806.

Sturm, I., Bach, S., Samek, W., and Müller, K.-R. (2016). Interpretable deep neural networks for single-trial EEG classification. *CoRR* abs/1604.0. doi: 10.1016/j.jneumeth.2016.10.008

Suk, H.-I., Lee, S.-W., and Shen, D. (2014). Hierarchical feature representation and multimodal fusion with deep learning for AD/MCI diagnosis. *NeuroImage* 101, 569–582. doi: 10.1016/j.neuroimage.2014.06.077

Thomas, A. W., Heekeren, H. R., Müller, K.-R., and Samek, W. (2018). Interpretable LSTMs for whole-brain neuroimaging analyses. *arXiv* arXiv:1810.09945.

Velayudhan, L., Proitsi, P., Westman, E., Muehlboeck, J.-S., Mecocci, P., Vellas, B., et al. (2013). Entorhinal cortex thickness predicts cognitive decline in Alzheimer's disease. *J. Alzheimer's Dis.* 33, 755–766. doi: 10.3233/JAD-2012-121408

Vieira, S., Pinaya, W. H., and Mechelli, A. (2017). Using deep learning to investigate the neuroimaging correlates of psychiatric and neurological disorders: methods and applications. *Neurosci. Biobehav. Rev.* 74, 58–75. doi: 10.1016/j.neubiorev.2017.01.002

Weiner, M. W., Veitch, D. P., Aisen, P. S., Beckett, L. A., Cairns, N. J., Green, R. C., et al. (2013). The Alzheimer's disease neuroimaging initiative: a review of papers published since its inception. *Alzheimer's Dement.* 9, e111–e194. doi: 10.1016/j.jalz.2013.05.1769

WHO (2017). *Dementia*. Available online at: https://www.who.int/news-room/fact-sheets/detail/dementia (accessed July 21, 2019).

Yang, C., Rangarajan, A., and Ranka, S. (2018). Visual explanations from deep 3D convolutional neural networks for Alzheimer's disease classification. *arXiv* arXiv:1803.02544.

Yosinski, J., Clune, J., Nguyen, A., Fuchs, T., and Lipson, H. (2015). Understanding neural networks through deep visualization. *arXiv* arXiv:1506.06579.

Zeiler, M., and Fergus, R. (2014). "Visualizing and understanding convolutional networks." in *Computer Vision–ECCV 2014*, volume 8689 of *Lecture Notes in Computer Science*, eds D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars (Zürich: Springer International Publishing), 818–833.

Zhang, X., Mormino, E. C., Sun, N., Sperling, R. A., Sabuncu, M. R., Yeo, B. T. T., et al. (2016). Bayesian model reveals latent atrophy factors with dissociable cognitive trajectories in Alzheimer's disease. *Proc. Natl. Acad. Sci. U. S. A.* 113, E6535–E6544. doi: 10.1073/pnas.1611073113