Visualizing evidence for Alzheimer’s disease in deep neural networks trained on structural MRI data

Moritz Böhle ¹, Fabian Eitel ¹, Martin Weygandt ² and Kerstin Ritter ¹,*, for the Alzheimer’s Disease Neuroimaging Initiative +

¹Charité – Universitätsmedizin Berlin, corporate member of Freie Universität Berlin, Humboldt-Universität zu Berlin, and Berlin Institute of Health; Department of Psychiatry and Psychotherapy, Bernstein Center for Computational Neuroscience; 10117 Berlin, Germany
²Charité – Universitätsmedizin Berlin, corporate member of Freie Universität Berlin, Humboldt-Universität zu Berlin, and Berlin Institute of Health; Excellence Cluster NeuroCure; 10117 Berlin, Germany.

+ Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf

Correspondence*:
Kerstin Ritter; Charitéplatz 1, 10117 Berlin, Germany
kerstin.ritter@charite.de

ABSTRACT

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 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. Thus, the highlighted areas can be interpreted as the ‘positive evidence’ used by the network for deciding whether an individual has AD. We find that this LRP-evidence indeed fulfills those expectations that one would have towards AD evidence: (1) it is very specific for individuals (“Why does this person have AD?”) with high inter-patient variability, (2) there is very little evidence for AD in healthy controls and (3) areas that exhibit a lot of evidence correlate well with what is known from literature. To quantify the latter, we compute size-corrected metrics of the summed evidence per brain area, e.g. the ‘evidence density’ or ‘evidence 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 hippocampus and amygdala. We conclude that LRP provides a powerful tool for assisting clinicians in finding evidence for AD (and potentially other diseases) in 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
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 grey 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 Tofghi, 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 Fig. 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. For example, the aforementioned methods of sensitivity analysis by Simonyan et al. (2013) or guided backpropagation by Springenberg et al. (2014) base the heatmaps on the gradient of the output with respect to the input. While the resulting heatmaps can be very informative, they do not allow for being interpreted as ‘evidence’. Instead, the susceptibility of the output to changes in the input are highlighted. Of course, the areas of high susceptibility might coincide with those areas on which the network bases its decision (“If the evidence is removed, the AD prediction will change.”), but they do not have to; in fact, the gradient heatmap might just point towards adversarial examples (Szegedy et al., 2013), i.e. slight alterations of input images that manage to fool the network. In medical applications, however, heatmaps that ‘explain the decision’ for the specific and individual input (e.g. the MRI image of a patient) are arguably of more value than heatmaps that tell you how you could ‘change the decision’.

Visualizing evidence for AD

Böhle et al.
A promising approach here 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 and can be interpreted as more specific evidence speaking for the one or the other class. 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 (ADNI). Based on the trained network in Rieke et al. (2018), we generated LRP heatmaps for each subject. 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. On an individual level, we analyzed the heatmap patterns of single subjects (‘relevance fingerprinting’) and looked into differences between AD patients. We show that the LRP heatmaps correlate more closely with the common conception of evidence and that this algorithm holds 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](http://www.adni-info.org).

For this study, we used the same data set as in [Rieke et al., 2018]. 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 labelled 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 946 individual scans (475 AD, 471 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). 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. 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 layers, 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 neighbouring 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).

For a detailed description of the CNN architecture and the training procedure, the reader is kindly referred to Rieke et al. (2018); the trained model can be downloaded from github. In short, the model 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 \( 3\times3\times3 \). Every convolutional layer is followed by batch normalization and max pooling with window sizes \( w \times w \times w \) (\( w = 3, 2, 3, 2 \)). The fully connected layers contain 128 and 64 units respectively. A final fully connected layer with 2 neurons and softmax activation serves as the network output, providing the class scores for HCs (first unit) and AD (second unit) respectively. The network was trained on a subset of 163 AD patients (399 images) and 121 HCs (395 images). For the visualization, the network was evaluated on the test set of 30 AD patients (76 images) and 30 HCs (99 images), on which it achieved a classification accuracy of 82%. The generalized classification performance was estimated via 5-fold cross-validation to be around \( 77 \pm 6\% \). While this accuracy is comparable to other studies, in which convolutional neural networks were used (Payan and Montana, 2015; Korolev et al., 2017), please note that the focus of the current work lies on the visualization of evidence used by the network and not on the state of the art classification performance. In recent work, higher generalized classification scores were achieved (e.g. Esmaeilzadeh et al. (2018)).

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 \( R_{i,l+1}^j \) of that node. Overall, the sum over the relevances of all nodes \( i \) contributing to neuron \( j \) in layer \( l \) must sum to \( R_{i,l+1}^j \), such that the total relevance per layer is conserved:

\[
\sum_i R_{i,l+1}^j = R_{l+1}^j
\]  

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. For the results in this work, the \( \beta \)-rule as described in a follow-up paper on LRP by Binder et al. (2016) has been used with \( \beta = 0.5 \). We chose this rule, as it allows for adjusting how much weight is put on positive contributions relative to inhibitory contributions that benefit the AD score. Given that we aim to produce heatmaps of positive evidence, with our choice of the \( \beta \)-parameter we placed more weight on positive contributions than on inhibitory contributions.

\[^{5}\text{https://github.com/jrieke/cnn-interpretability}\]  
\[^{6}\text{The split of the data is performed on the level of patients instead of images to prevent the network from seeing images of the same patient during training and testing.}\]
The relevance propagation rule can be written as

\[ R_{i,l,l+1}^{j} = \left( 1 + \beta \frac{z_{ij}^+}{z_j^+} \right) R_{j,l+1}^j - \beta \frac{z_{ij}^-}{z_j^-} R_{j,l+1}^j. \]  

(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 \( l \),

\[ z_{ij}^+/^- = \sum_i z_{ij}^+/^- \], such that the relevance is conserved from layer \( l + 1 \) to layer \( l \).

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.

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; in this work we relied on the GB implementation by Rieke et al. (2018). 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. 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 and in the following chapters, we refer to this as ‘evidence’ for AD. It is very important to note that this ‘evidence’ does not necessarily relate to any ‘true evidence’ in the sense of biomarkers and the quality of the evidence highly depends on the data, the relevance-propagation rule, and the strength of the trained network. Any heatmap method can only try to approximate the influence that any specific input voxel had on the final result and the LRP method is no exception. Nevertheless, if the approximation is strong enough and LRP succeeds in breaking down the relevance to the level of voxels in a meaningful way, one can hope that the highlighted ‘evidence’ indicates the presence of a biomarker that the network has learned to correlate with AD. In section 4, several ways in which the quality of the heatmaps can be assessed, are delineated.

All LRP heatmaps show the average evidence for the same class (AD), such that they can be compared on the same scale (evidence for AD). As not all HCs were given an AD score of 0%, there will be evidence for AD in the controls, too. As an example of such heatmaps, a sagittal slice of the average AD heatmap is presented in Fig. 2.

2.5 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.

https://www.github.com/moboehle
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 Fig. 3. 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.

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 evidence 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). While low values over large areas might be due to statistical fluctuations in the data, clusters of evidence (LRP) or susceptibility (GB) in a very confined area could be indicative of the presence of certain biomarkers for AD.

Gain – ratio of values with respect to the average HC. Lastly, it is important to note that HCs are not ‘evidence-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. On average, HCs were assigned a class score of 22% over the test set. Thus, as an additional metric, we will look at the ‘gain’ in evidence (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 Figs. 3-5, we depict (1) sum of AD importance per area (Fig. 3), (2) size-normalized AD importance per area (Fig. 4) and (3) gain per area (Fig. 5), separately for GB and LRP. For better visibility, only the top 25 most ‘important’ areas are shown for each metric and visualization method respectively. In section 3.1, we compare GB and LRP with respect to the importance metrics and their average heatmaps for different sets of data (AD, HC, true positives, false positives etc.). As the LRP algorithm allows for more interpretable individual heatmaps, we will present and discuss the LRP heatmaps of two individual patients in section 3.2.

3.1 GB vs. LRP

In Fig. 3, we show the sum of AD importance per area, separately for LRP and GB. Although this metric seems to be strongly 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 the cerebellum is even higher for HC than for AD. This exemplifies well that the heatmaps for GB cannot be interpreted as positive evidence for AD, 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 shows a very strong correlation to the size of the respective brain area, the following metrics, in which we controlled for the brain area size, are better interpretable.

In Fig. 4, the total sum of importance is normalized by the size of the respective brain area. Several important features, which show the potential strength of LRP as a diagnostic tool, are well captured in these visualizations. 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. More importantly, 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 (Park et al., 2017; Ferreira et al., 2017). In contrast, for HCs the evidence 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 Fig. 4; for these patients the individual heatmaps will be compared in section 3.2 and Fig. 6.

In Fig. 5, 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 the amygdala and the hippocampus. Additionally to showing the distributions in
Fig. 5, the average heatmaps for the distinct classification cases (true positives, false positives, etc.) are shown in Fig. 8. In particular, the false positives lead to an interesting insight: For LRP, the false positives exhibit less evidence 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 evidence was less compelling than for true AD patients, leading to lower AD class scores. Since in LRP the evidence scales with the class score – the class score is redistributed onto the input voxels – the fact that the evidence is weaker is built into the LRP algorithm. For GB on the other hand, the false positives seem to exhibit the highest gradient values of all cases. This exemplifies well what GB truly measures: in the case of false positives, 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 not necessarily represent areas with true evidence for AD, but instead might just lie on the shortest path for leaving the data manifold the network model was trained on. Outside of this manifold, the decisions of neural networks are unpredictable (see e.g. adversarial examples, (Szegedy et al., 2013)). In contrast to this, for true positives in GB we see a clear focus on the hippocampus (top row in the Figure), which coincides with what is highlighted by LRP. In this case, the highlighted area in GB might be more meaningful and correspond to actual evidence; the gradient method might show in this case that if this ‘positive evidence’ were to be changed, the outcome of the network would strongly change, too.

3.2 Individual heatmaps - fingerprinting

Since the LRP algorithm highlights positive contributions to the class score of AD, the resulting heatmaps correspond more closely to our common conception of evidence. For each individual patient the heatmap takes into account the actual filter activations of the current sample throughout the layers of the model and thus the heatmaps might serve as ‘individual fingerprints’ in a diagnostic tool. In Fig. 6, we show several slices of the evidence heatmaps for two patients in order to highlight the diversity in those heatmaps. The two patients were selected as those with the highest Euclidean distance in the evidence-density space between each other among those patients that were strongly classified as AD (class score > 95%; individual trajectories of region-wise evidence are shown in Fig. 4). It can be seen that the areas, which mainly contributed to the network decision, are very different for the two patients. For one patient, the class score of the network is heavily influenced by the hippocampus, while for the second patient, the area around the planum temporale and the transversal temporal gyrus seem to be the most affected.

4 DISCUSSION

In this study, we compared two methods, namely GB and LRP, for explaining individual CNN decisions in AD classification. After training a CNN to separate AD patients and HCs based on structural MRI data, we produced individual heatmaps for the test subjects. These heatmaps indicate the importance for each voxel for the respective classification decision. We analyzed the data separately for AD patients and HCs, and for different classification outcomes (true positive, false positive etc.). The relevance of brain regions contained in the Neuromorphometrics atlas were 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 can be interpreted as AD evidence as they provide (1) high specificity for individuals (2) little evidence for AD in HCs and (3) areas that exhibit a lot of evidence correlate well with what is known from literature. These evidence 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 is 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 ‘evidence hot-spots’ across all AD patients. Furthermore, the size-corrected metrics (‘evidence density’ and ‘evidence 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. Judging from the gain-metric, especially the hippocampus and the amygdala seem to carry a lot of information on the state of the patient. This surely gives reason to hope 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 and GB

We quantitatively evaluated the heatmaps, obtained by either GB or LRP, towards different brain areas according to the Neuromorphometrics atlas (Bakker et al., 2015) by summarizing the importance (AD evidence 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, temporal pole, middle temporal gyrus, parahippocampal gyrus, and entorhinal cortex (Du et al., 2001; Desikan et al., 2009; Frisoni et al., 2010; Weiner et al., 2013; Velayudhan et al., 2013; Klein-Koerkamp et al., 2014; Long et al., 2017) as well as cortical regions including basal forebrain, insula and parietal regions (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 grey 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 gain (defined as the ratio between the parcellated heatmaps of true AD patients and HCs) is much higher for LRP than GB meaning that the heatmap differences between AD patients and HCs are less pronounced in GB. This supports the notion that LRP heatmaps can be interpreted as evidence heatmaps, whereas GB emphasizes areas which the network more generally is sensitive to. Notably, the gain of AD evidence in LRP was highest for two key structures in AD, hippocampus and amygdala. This shows quite impressively that the proposed CNN actually captures information from relevant brain areas. Regarding other structures found to be important in our network, it might be interesting to see if also other neural networks find evidence 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 LRP for fingerprinting

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 Fig. 6. 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. It might be interesting to investigate whether we can confirm these results via clustering and if LRP fingerprints correlate with clinical markers and neuropsychological test scores such as sub domains of the Alzheimer’s Disease Assessment Scale (ADAS). Moreover, our metrics should be evaluated in patients with mild cognitive impairment (MCI). By quantifying the amount of evidence for AD in these MCI patients, the progression of the disease might be tractable and early diagnosis of AD would be facilitated.

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. These studies either visualized filters and activations of the first or last layer (Sarraf and Tofighi, 2016; Ding et al., 2019; Lu et al., 2018) 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. Additionally, they did not compare different visualization methods or looked for inter-individual
differences. Another study 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 LSTM 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 very 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.

The first limitation of LRP is closely related to one of its strengths. There is no ground truth for the heatmaps and while our study seems to indicate that the LRP heatmaps correlate well with a common conception of evidence, they are only an approximation to what influences the classifier in its decisions. Generally, heatmaps are very sensitive to the specific algorithms (and its parameters, e.g. the $\beta$-parameter for LRP) used to produce them. Thus, heatmaps can be easily misinterpreted, especially by practitioners who are not familiar with the underlying algorithms. Most commonly, heatmaps are qualitatively evaluated based on prior knowledge (e.g. hippocampus is known to be strongly affected in AD, therefore it seems reasonable to find evidence there), 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. Given that in the specific case of heatmaps for MR images the input space is highly structured (all brains are registered to a common space), additional ways for assessing the quality of the evidence exist. On the one hand, it allows for correlating the importance of areas (according to the heatmaps) with literature. On the other hand, important structures such as the hippocampus can easily be removed from the input images and a new network can be trained on brains in which a presumably important brain area is completely removed. If the drop in accuracy correlates strongly with the supposed importance (e.g. according to our metrics), the assumption that LRP indeed shows evidence could be corroborated.

Second, 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.

Third, heatmaps strongly depend on the 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 classification decisions in a CNN trained to separate AD patients and HCs based on structural MRI data. In contrast to GB, LRP heatmaps can be interpreted as providing individual AD evidence (“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., ‘evidence density’ and ‘evidence 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 can benefit from further optimizing the network architecture 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 might try to identify subtypes of AD or MCI. Furthermore, we are convinced that our framework might also be very useful for other disease classification studies in helping to understand individual network decisions.

CONFLICT OF INTEREST STATEMENT

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

AUTHOR CONTRIBUTIONS

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

FUNDING

We acknowledge support from the German Research Foundation (DFG, 389563835), the Manfred and Ursula-Müller Stiftung and Charité – Universitätsmedizin Berlin (Rahel-Hirsch scholarship and Open Access Publication Fund).

ACKNOWLEDGMENTS

Data collection and sharing for this project was funded by the Alzheimer’s Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer’s Association; Alzheimer’s Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer’s Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

DATA AVAILABILITY STATEMENT

The ADNI data set is for researchers publicly available at [http://adni.loni.usc.edu/](http://adni.loni.usc.edu/). The code is available at [https://www.github.com/moboehle](https://www.github.com/moboehle).
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 Transactions on audio, speech, and language processing 22, 1533–1545.

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. 9525–9536.

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, 1–46. 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.

Binder, A., Montavon, G., Lapuschkin, S., Müller, K.-R., and Samek, W. (2016). Layer-wise relevance propagation for neural networks with local renormalization layers. In International Conference on Artificial Neural Networks (Springer), 63–71.

Bondi, M. W., Edmonds, E. C., and Salmon, D. P. (2017). Alzheimer’s disease: Past, present, and future. Journal of the International Neuropsychological Society 23, 818–831.

Braak, H. and Braak, E. (1991). Neuropathological stageing of Alzheimer-related changes. Acta Neuropathologica 82, 239–259. doi:10.1007/BF00308809.

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. Frontiers in Neuroinformatics 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. Journal of Neurology, Neurosurgery & Psychiatry 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 (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. Scientific reports 7, 46263.

Frisoni, G. B., Fox, N. C., Jack Jr, C. R., Scheltens, P., and Thompson, P. M. (2010). The clinical use of structural MRI in Alzheimer disease. Nature Reviews Neurology 6, 67.

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), eds. S. Dasgupta and D. Mcallester (JMLR Workshop and Conference Proceedings), vol. 28, 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. Frontiers in aging neuroscience 6, 20.

Kilimann, I., Hausner, L., Fellgiebel, A., Filipp, M., Würdemann, T. J., Heinsen, H., et al. (2017). Parallel atrophy of cortex and basal forebrain cholinergic system in mild cognitive impairment. Cerebral Cortex 27, 1841–1848.

Klein-Koerkamp, Y., Heckemann, R., Ramdeon, K., Moreaud, O., Keignart, S., Krainik, A., et al. (2014). Amygdalar Atrophy in Early Alzheimer’s Disease. Current Alzheimer Research 11, 239–252. doi:10.2174/156720501166140131123653.

Klöppel, S., Stonington, C. M., Chu, C., Draganski, B., Scabill, R. L., Rohrer, J. D., et al. (2008). Automatic classification of MR scans in Alzheimer’s disease. Brain 131, 681–689.

Korolev, S., Safullin, 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) (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 (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. 2912–2920

LeCun, Y. and Bengio, Y. (1995). Convolutional networks for images, speech, and time series. The handbook of brain theory and neural networks 3361

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. Nature 521, 436

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. Scientific reports 8, 11258

Lipton, Z. C. (2018). The Mythos of Model Interpretability. Queue 16, 30:31–30:57. doi:10.1145/3236386.3241340

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. Medical image analysis 42, 60–88

Liu, M., Cheng, D., Wang, K., and Wang, Y. (2018). Multi-Modality Cascaded Convolutional Neural Networks for Alzheimer’s Disease Diagnosis. Neuroinformatics , 1–14: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. 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

Lu, D., Popuri, K., Ding, G. W., Balachandar, R., Beg, M. F., and Alzheimer’s Disease Neuroimaging Initiative, A. D. N. (2018). Multimodal and Multiscale Deep Neural Networks for the Early Diagnosis of Alzheimer’s Disease using structural MR and FDG-PET images. Scientific reports 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 Processing 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. doi:10.1212/WNL.43.11.2412-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. The Lancet Neurology 10, 785–796. doi:10.1016/S1474-4422(11)70156-9

Noh, Y., Jeon, S., Lée, J. M., Seo, S. W., Kim, G. H., Cho, H., et al. (2014). Anatomical heterogeneity of Alzheimer disease based on cortical thickness on MRIs. Neurology 83, 1936–1944

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. Neuroscience & Biobehavioral Reviews 36, 1140–1152

Park, J.-Y., Na, H. K., Kim, S., Kim, H. J., Seo, S. W., et al. (2017). Robust identification of Alzheimer’s disease subtypes based on cortical atrophy patterns. Scientific reports 7, 43270

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. Journal of Neurology, Neurosurgery & Psychiatry 84, 556–561. doi:10.1136/jnnp-2012-303299

Rathore, S., Habes, M., Iftiikhar, 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

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 (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. Journal of Alzheimer’s Disease , 1319–1331

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 & Dementia: Diagnosis, Assessment & Disease Monitoring 1, 206–215
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 preprint 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

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

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

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

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., et al. (2013). Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*

Thomas, A. W., Heekeren, H. R., Müller, K.-R., and Samek, W. (2018). Interpretable LSTMs For Whole-Brain Neuroimaging Analyses. *arXiv preprint 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. *Journal of Alzheimer’s Disease* 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. *Neuroscience & Biobehavioral Reviews* 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 & Dementia* 9, e111–e194

WHO (2017). Dementia. Available at: https://www.who.int/news-room/fact-sheets/detail/dementia

Yang, C., Rangarajan, A., and Ranka, S. (2018). Visual Explanations From Deep 3D Convolutional Neural Networks for Alzheimer’s Disease Classification. *arXiv preprint arXiv:1803.02544*

Yosinski, J., Clune, J., Nguyen, A., Fuchs, T., and Lipson, H. (2015). Understanding Neural Networks Through Deep Visualization. *arXiv preprint arXiv:1506.06579*

Zeiler, M. and Fergus, R. (2014). Visualizing and Understanding Convolutional Networks. In *Computer Vision – ECCV 2014*, eds. D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars (Springer International Publishing), vol. 8689 of *Lecture Notes in Computer Science*. 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. *Proceedings of the National Academy of Sciences of the United States of America* 113, E6535–E6544. doi:10.1073/pnas.1611073113
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 evidence 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.
Figure 2. Evidence (dark to bright: low to high evidence) heatmap in a saggital brain slice obtained as an average over all AD patients via the LRP algorithm. This indicates that areas of interest to the network include the hippocampal area (horizontal bright strip in the center) and the insula (vertical bright strip to the right).
Figure 3. Absolute sum of evidence (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 highly correlated with the size of the brain areas.
Figure 4. Size-normalized evidence (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 evidence distributions for LRP; the patients were selected as those with the highest Euclidean distance in the evidence-density space between each other among those patients that were strongly classified as AD (class score >95%).
Figure 5. Gain of evidence (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 evidence (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 true evidence, 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.
Figure 6. Three brain slices are shown for patient A and patient B, whose individual slopes in evidence density have been shown in Figure 4. The first slice (top: 60) was selected to contain the largest cross-section of the hippocampus (green). The other two slices (middle: 86, bottom: 91) were selected to contain the largest cross-sections of those five areas, in which the two patients have the greatest difference in evidence density. These areas are in decreasing order the planum temporale (red), caudate (dark blue, second and third row), the transversal temporal gyrus (cyan, second and third row), the superior temporal gyrus (black), and the amygdala (light blue, first row). 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 7. Average heatmaps for AD patients and healthy controls (HCs) are shown separately for LRP (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 8. For both visualization methods – LRP (left) and GB (right) – 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. For each heatmap, the color-coding is the same as in Fig. 7 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.