Received: 29 March 2021 | Revised: 6 September 2021 | Accepted: 9 September 2021

DOI: 10.1002/ima.22657

WILEY

RESEARCH ARTICLE

Interpretability of deep neural networks used for the diagnosis of Alzheimer's disease

Tomáš Pohl | **Marek Jakab** | **Wanda Benesova**

Institute of Computer Engineering and Applied Informatics, Faculty of Informatics and Information Technologies STU, Bratislava, Slovakia

Correspondence
Tomáš Pohl, Institute of Computer Engineering and Applied Informatics, Faculty of Informatics and Information Technologies STU in Bratislava, Ilkovičova 2, 842 16 Bratislava, Slovakia.
Email: tomas.pohl@gmail.com

Funding information
Canadian Institutes of Health Research; Transition Therapeutics; Takeda Pharmaceutical Company; Servier; Piramal Imaging; Pfizer Inc.; Novartis Pharmaceuticals Corporation; NeuroRx Research; Neurotrack Technologies; Meso Scale Diagnostics, LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; IXICO Ltd.; GE Healthcare; Fujirebio; Genentech, Inc.; F. Hoffmann-La Roche Ltd; Eli Lilly and Company; EuroImmun; Elan Pharmaceuticals, Inc.; Eisai Inc.; Cogstate; CereSpir, Inc.; Bristol-Myers Squibb Company; Biogen; Araclon Biotech; BioClinica, Inc.; Alzheimer's Drug Discovery Foundation; AbbVie, Alzheimer's Association; National Institute of Biomedical Imaging and Bioengineering; National Institute on Aging; Department of Defense (DOD), Grant/Award Number: W81XWH-12-2-0012; National Institutes of Health, Grant/Award Number: U01 AG024904; Alzheimer's Disease Neuroimaging Initiative (ADNI); STU Grant scheme for Support of Young Researchers

Abstract

Alzheimer's disease (AD) is a chronic brain disorder and is the most common cause of dementia. Patients suffering from AD experience memory loss, confusion, and other cognitive and behavioral complications. As the disease progresses, these symptoms become severe enough to interfere with the patient's daily life. Since AD is an irreversible disease and existing treatments can only slow down its progress, early diagnosis of AD is a key moment in fighting this disease. In this article, we propose a novel approach for diagnosing AD via deep neural networks from magnetic resonance imaging images. Additionally, we propose three new propagation rules for the layer-wise relevance propagation (LRP) method, which is a method used for visualizing evidence in deep neural networks to obtain a better understanding of the network's behavior. We also propose various rule configurations for the LRP to achieve better interpretability of the network. Our proposed classification method achieves a 92% accuracy when classifying AD versus healthy controls, which is comparable to state-of-the-art approaches and could potentially aid doctors in AD diagnosis and reduce the occurrence of human error. Our proposed visualization approaches also show improvements in evidence visualization, which helps the spread of computer-aided diagnosis in the medical domain by eliminating the "black-box" nature of the neural networks.

KEYWORDS

Alzheimer's disease, deep neural networks, interpretability, layer-wise relevance propagation, magnetic resonance imaging

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.

1 | INTRODUCTION

Alzheimer's disease (AD) is a progressive and irreversible chronic brain disorder and is the most common cause of dementia. People suffering from AD, among other symptoms, may experience memory loss, confusion and disorientation, personality issues, and, ultimately, the loss of bodily functions.[1] According to a recent study taken in 2018, 50 million people suffer from dementia worldwide, from which 50%–60% are cases of AD, and by 2050 this number is expected to triple to 152 million.

There are two major contributors to the formation of this disease, namely *amyloid plaques* and *neurofibrillary tangles*, which prevent the communication between neurons and lead to their death.[2] AD is, therefore, mainly characterized by a loss of large number of neurons in the brain, which is called brain atrophy.

Due to the damaged neurons that die throughout the brain, certain regions of the brain start to shrink. The first signs of AD appear in the entorhinal cortex and in the hippocampus.[3] As the disease progresses, other parts of the brain begin to shrink as well. This brain tissue shrinkage is visible in magnetic resonance imaging (MRI) scans of the brain, which is therefore often used to diagnose and monitor the progress of AD. Although, in recent years, new and more precise techniques have been developed to provide a better aid in AD diagnosis, namely amyloid and tau imaging using positron emission tomography (PET),[4] their limited availability and high cost prevent their widespread usage, hence MRI remains the major imaging modality used for AD diagnosis.[5]

In past years, research studies have been actively using machine learning approaches to process medical data to aid doctors in diagnosing AD. In 2015, Payan and Montana[6] compared the performance of 2D and 3D convolutional networks, for which they used a model that combines sparse autoencoder and 2D, resp. 3D, architecture. They evaluated their approach on an MRI dataset and concluded that although in the case of the AD versus healthy controls (HC) classification the accuracy was identical (95.39%), when they introduced the mild cognitive impairment (MCI) class, the 3D-CNN (89.47%) clearly outperformed the 2D-CNN (85.83%).

Liu et al published two papers in 2018 where they used deep learning for AD diagnosis. In the first paper,[7] the authors used a combination of convolutional and recurrent neural networks to process PET images and achieved an accuracy of 91.2% for AD versus HC classification. In their second paper,[8] landmark detection and 3D-CNN were used to process MRI scans. The achieved accuracy for AD versus HC binary classification was 91.09%. In 2020, Liu et al[9] proposed a multi-model deep CNN for automatic hippocampus segmentation and classification in AD. With this approach, they achieved an accuracy of 88.9% in classifying AD versus HC.

In one of the latest research, Mehmood et al[10] utilized layer-wise transfer learning and tissue segmentation for early stage AD diagnosis. They achieved 98.73% accuracy when classifying AD versus HC and 83.72% accuracy when classifying early versus late MCI patients.

For more information about state-of-the-art research in the domain of medical data processing for the diagnosis of AD, please see the recent review articles by Jo et al,[11] Ebrahimighahnavieh et al,[12] or Tanveer et al.[13]

1.1 | Interpretability of neural networks

Although CNNs can achieve good results and can aid the doctors in diagnosing AD, their inability to provide information about how and why they arrive at a certain decision limits (or even prevents) their usage in certain domains. Such domain is, for example, the medical domain, where a wrong diagnosis can almost always result in danger to human life.

Additional motivation behind understanding the decisions of machine learning models is to expose "Clever Hans" predictors[14] or to identify new useful features, but in some cases, explanations of AI systems are even part of the legislation.[15]

To overcome this "black-box" nature of CNNs, various methods have been proposed in recent years. In 2013, Simonyan et al[16] introduced the *sensitivity analysis*, which determines the contribution of each input feature (e.g., pixels) to the output via gradients. A modification of the sensitivity analysis is the *guided backpropagation* (GB),[17] which only considers gradients that have positive error signal. Since these methods tend to produce noisy outputs, Shrikumar et al[18] proposed the *gradient* $\times$ *input* method, which in general produces more focused interpretations.

Although these gradient-based interpretation techniques are scalable and easy to implement, it has been argued that they only measure the susceptibility of the output to changes in the input, and therefore the features identified as relevant might not align with features that the network bases its decision on.[19]

Another group of interpretation methods is the perturbation-based methods, which calculate the relevance of the input features by comparing the original image's output with a masked (perturbed) image's output. Such method is the *occlusion sensitivity* introduced by Zeiler and Fergus.[20] This approach systematically occludes different parts of the input image with usually black or gray patches while monitoring the output. Although the perturbation-based methods can be easily

applied to any model that has an evaluation function, they are very computation-intensive.

In 2018, Rieke et al[21] trained a 3D-CNN to diagnose AD based on MRI scans, while they also compared four different interpretation methods, specifically sensitivity analysis, GB, occlusion where they systematically occluded the image with black patches of size $40 \times 40 \times 40$, and brain area occlusion where they occluded an entire brain area. Although each method identified brain regions that are known biomarkers of AD, they also stated that gradient-based methods are better than occlusion-based methods in cases where the relevance is presumably distributed across the input image, since the occlusion-based methods are unable to capture larger areas (e.g., cortex, which is a known biomarker of AD). The experiments were conducted on a model that had an accuracy of 77% for AD versus HC.

A third group of interpretation methods is the relevance-based methods. Such method is the *layer-wise relevance propagation* (LRP) introduced by Bach et al,[22] which explains the classification predictions via pixel-wise decomposition of nonlinear classifiers. LRP calculates the relevance of each input feature by back-propagating the neural network's prediction by means of purposely designed local propagation rules.[23] This approach is discussed in more detail in section 2.

In 2019, Böhle et al[19] used LRP to explain CNN decisions during AD diagnosis based on structural MRI scans and compared it to the GB. The authors concluded that the LRP method is superior to the gradient-based GB method in the case of AD diagnosis, as LRP is more image-specific, can better distinguish between AD and HC subjects, and shows the most AD evidence in true positive AD samples, while the GB showed the most AD evidence in false classifications.

In this article, we propose a novel 3D-CNN architecture to classify AD based on MRI structural scans while we also propose various modifications to the LRP interpretation method to explain the decision of the proposed CNN. Specifically, we propose three new LRP propagation rules—$input \times w^2$ (section 2.3.1), w-log (section 2.3.2), and $\sqrt{w}$ (section 2.3.3), and the usage of two existing LRP approaches (sections 2.3.4 and 2.3.5), which, according to the best of our knowledge, have not yet been applied for the classification of MRI scans. Additionally, we compare our proposed interpretation methods to existing ones both qualitatively and quantitatively.

The article is structured as follows. Section 2 describes the used dataset, the proposed 3D-CNN architecture as well as the proposed LRP rules alongside various LRP approaches. This section also details the used evaluation methods. Section 3 presents the achieved classification results compared with state-of-the-art approaches, while also evaluating the proposed interpretation approaches from both qualitative and quantitative point of view. Finally, section 4 concludes the main findings of this paper.

2 | MATERIALS AND METHODS

2.1 | Dataset

The used dataset in this work is part of the TADPOLE challenge[24] and was obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI)*, whose primary goal since 2003 has been to test whether various modalities, clinical, and neuropsychological assessments can be combined to measure the progression of AD. The MRI scans in our dataset were acquired via three-dimensional, T1-weighted, gradient-echo sequence, called magnetization-prepared rapid gradient-echo (MP-RAGE).

The dataset contains a total of 7426 entries of 1447 unique subjects who can be assigned three possible diagnoses: HC, MCI, and AD. Since a considerable part of our research effort has been put into the interpretability of neural networks, the MCI group has been filtered out to ensure a more reliable evaluation and comparison of the proposed visualization methods, and therefore further research was carried out on a binary classification model. Additionally, each subject was assigned precisely one class. This means that subjects with multiple diagnoses were assigned the worst one (e.g., if a patient had MCI and AD diagnosis, he was assigned the AD class). The final dataset, therefore, contained a total of 3634 MRI scans obtained from 969 subjects: 2033 HC scans (462 subjects) and 1601 AD scans (507 subjects). The demographic details are summarized in Table 1.

The downloaded dataset has been preprocessed with a standard ADNI pipeline. For MRI scans, this includes correction for gradient non-linearity, B1 non-uniformity correction, and peak sharpening.[24] Additionally, the dataset used by us has also been skull stripped, adjusted to the same orientation, re-sampled to isotropic resolution, and all the scans have been resized to have identical dimensions. Finally, the inconsistencies between different scanners were addressed by standardizing the dataset.

TABLE 1 Demographic details of the subjects

	n	Gender (M/F)	Age	MMSE
HC	462	230/232	73.9 ± 6.2	29.0 ± 0.9
MCI	478	276/202	73.1 ± 7.6	27.3 ± 2.2
AD	507	287/220	74.2 ± 7.6	22.2 ± 3.4

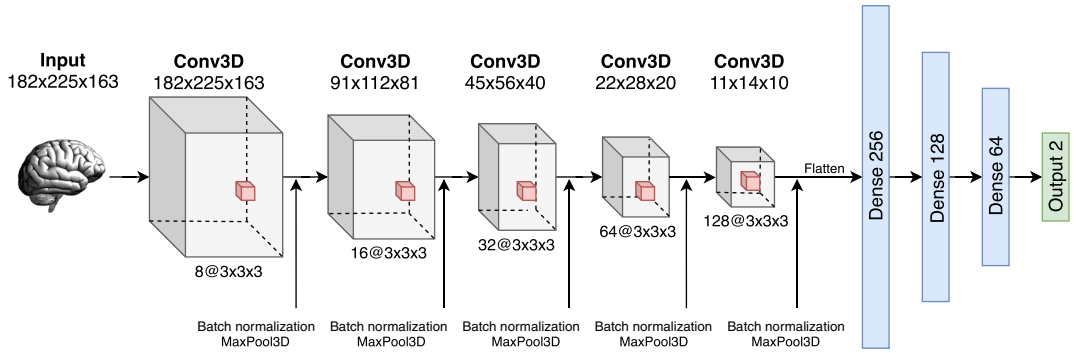

FIGURE 1 Architecture of the proposed ADNet model. Each convolutional layer is followed by a batch normalization and a max-pooling layer

2.2 | Proposed CNN architecture

For the binary classification of AD versus HC, a 3D-CNN was proposed, as according to recent studies analyzed in section 1, the 3D-CNNs tend to outperform 2D-CNNs. 3D-CNNs are also useful in AD diagnosis where the three-dimensional nature of the model is able to capture various spatial patterns and structures.

The proposed neural network architecture, which bears the name ADNet, has five convolutional blocks with 8/16/32/64/128 feature maps, three fully-connected layers with 256, 128, and 64 neurons, and an output layer with two neurons and softmax activation. All convolutional and dense layers, except the output layer, utilize the ReLU activation function. The size of the max-pooling windows is set to 2 on all max-pooling layers. The used Adam optimizer has a learning rate of 1×10^{-4}. As the loss function, the categorical cross-entropy was used. The architecture of the model is illustrated in Figure 1, and the key parameters used during training are summarized in Table 2.

TABLE 2 Key model parameters used during model training

Learning rate	Activation function	Batch size	Dropout rate	Total epochs
1×10^{-4}	ReLU	5	0.8	20

2.3 | Proposed evidence visualization methods

As mentioned in section 1, we also propose modifications to the existing relevance-based LRP method which calculates the relevance of each input feature by back-propagating the neural network's prediction by means of purposely designed local propagation rules.[23] The basic rule of relevance score propagation between two consecutive layers of the neural network is defined as:

$$R_j = \sum_k \frac{z_{jk}}{\sum_j z_{jk}} R_k, \tag{1}$$

where j and k are neurons at two consecutive layers, $(R_k)_k$ are the propagated relevance scores, and z_{jk} is the extent to which neuron j has contributed to neuron k.[23] For a comprehensive list of existing propagation rules, please refer to the overview article by Montavon et al.[23]

LRP can be divided into two categories: uniform LRP and composite LRP. The difference between these two approaches is that while the uniform LRP approach uses a single rule for all the layers, the composite LRP utilizes multiple rules for a single interpretation. According to Montavon et al,[23] composite LRP delivers an explanation that is more faithful and understandable than uniform LRP.

Additionally, LRP, in contrast to gradient-based or perturbation-based methods, can also differentiate between positive and negative evidence. Positive evidence are features that support the outcome of the predicted class, while negative evidence are features that support the outcome of the opposite class(es). This property can be very powerful in the case of AD diagnosis, as the CNN interpretation will be able to show which brain regions support the outcome of AD and which the outcome of HC.

The implementation of the proposed LRP modifications can be found at our GitHub repository†, which is based on the existing iNNvestigate library.[25]

2.3.1 | The *input* $\times$ w^2 rule

One of the existing propagation rules is the w^2-*rule* (used mainly for the first layer), which is only indirectly dependent on the input data through the influence of $R^{(l)}$. We propose a formula in which we include the influence of the input data directly. The proposed modification, which we named as the *input* $\times$ w^2—rule, is defined as follows:

$$R_i = \sum_j \frac{x_i w_{ij}^2}{\sum_i x_i w_{ij}^2} R_j. \quad (2)$$

With this modified formula, we expect the final heatmap to be more dependent on the input and, therefore, more focused and less noisy. Similarly to the original w^2-rule, this modification is intended to be used on the input layer.

2.3.2 | The *w-log* rule

The next modification we propose is also an alteration of the w^2-rule and is intended for the first layer of the network. Instead of squaring the weights, we propose to apply the logarithmic function in order to better emphasize the negative evidence in the visualization.

This approach of better emphasizing the negative evidence is especially advantageous in the medical domain where the consequences of a bad classification are crucial. The proposed *w-log* rule takes a pessimistic standpoint and artificially enhances the features that support the outcome of the opposite class(es), which makes it harder to miss signs contradicting our decision.

The formula for this proposed *w-log* rule is defined as:

$$R_i = \sum_j \frac{\log_{10}(w_{ij})}{\sum_i \log_{10}(w_{ij})} R_j. \quad (3)$$

In order to preserve the positive relevance as positive and negative relevance as negative, one must also shift the logarithmic function one unit to the left. Another caveat of this rule is that it is only defined on the $(0, \inf)$ interval which becomes $(-1, \inf)$ after shifting the function one unit to the left.

2.3.3 | The *w-sqrt* rule

The additional new rule we propose is similar to those described above: we modify the w^2-rule by taking the square root of the weights rather than their square. With this approach, we intend to better emphasize the features that are not the most relevant but still contribute to the outcome of the classification.

The proposed $\sqrt{w}$ rule is useful in situations where the classification outcome is uncertain and one might want to better highlight the less significant contributions while keeping the order of the features from the relevant point of view intact. This rule, which is defined only on the positive interval and is advised to be used on the input layer, is given by the following formula:

$$R_i = \sum_j \frac{\sqrt{w_{ij}}}{\sum_i \sqrt{w_{ij}}} R_j. \quad (4)$$

2.3.4 | Composite LRP

Besides proposing various propagation rules, we also propose to use composite LRP besides uniform LRP. Although composite LRP is not a novel approach proposed by us, to the best of our knowledge, it is applied for the first time for an MRI classification problem.

While proposing various composite LRP configurations, we took into consideration the work by Montavon et al,[23] which describes which LRP rule is beneficial to use for which layer type. Our proposed composite LRP approaches, therefore, follow the following conventions:

- Upper layers: LRP-0 and LRP-ϵ
- Middle layers: LRP-ϵ and LRP-$\alpha\beta$
- Lower layers: LRP-$\alpha\beta$ and Flat
- First layer: w^2, Bounded,[26] *input* $\times$ w^2 (our), *w-log* (our), and $\sqrt{w}$ (our)

2.3.5 | Top layer modification

Montavon et al[23] also showed that if one wants to obtain explanations that contain negative evidence, it is beneficial to replace the $z_c = \sum_{0,k} a_k w_{kc}$ score, which is linked to the predicted class probability via the softmax function $P(\omega_c) = \exp(z_c)/\sum_{c'} \exp(z_{c'})$, with the following score: $\eta_c = \log[P(\omega_c)/(1 - P(\omega_c))]$.

This top layer modification can be expressed with the following sequence of layers[23]:

$$z_{c,c'} = \sum_{0,k} a_k (w_{kc} - w_{kc'}), \quad (5)$$

$$\eta_c = -\log \sum_{c' \neq c} \exp(-z_{c,c'}). \quad (6)$$

The first layer calculates the log-probability ratios, while the second layer performs a reverse log-sum-exp pooling over these ratios. To propagate the relevance through this pooling layer, a min-take-most strategy is advised to be used. Such strategy was proposed by Kauffmann et al[27]:

$$R_{j\leftarrow k} = \frac{\exp(-a_j)}{\sum_j \exp(-a_j)} \cdot R_k. \quad (7)$$

2.4 | Evaluation of the proposed evidence visualization methods

Although humans are able to intuitively assess the quality of the interpretations (heatmaps in our case) by matching it to known biomarkers, domain knowledge, and experience, they are unable to evaluate which input features are the most relevant for the classifier. In order to quantitatively evaluate the heatmap, one needs to define meaningful measures.

2.4.1 | Evaluation via pixel-flipping

Samek et al[28] and Binder et al[29] used an evaluation method in which they perturbed the input space and observed its impact on the prediction. Since the heatmap is an array of pixel-wise scores that indicate to what extent is a given pixel relevant for a specific classification decision, we can replace the most relevant ones with noise and measure the change in the output. To put it more formally, a pixel p is considered highly relevant for a given classification score $f(x)$ and a given image x, if replacing it with noise and classifying this modified image $\bar{x}_p$, the classification score $f(\bar{x}_p)$ will strongly decrease.

To evaluate our proposed methods with this approach, we can order the input features (voxels in our case) based on their relevance score, where the most relevant one will be in the first position, and start replacing the highest ranking voxels with noise while monitoring the output. The obtained results can be compared either to random flipping, during which we replace randomly selected voxels, or to minimum flipping, during which we start replacing the voxels with the lowest relevance score.

2.4.2 | Evaluation via Atlas-based importance metrics

Another quantitative evaluation metric is the Atlas-based importance metrics used by Böhle et al.[19] In this approach, we measure either the total relevance in each region of the brain or the size-normalized relevance (i.e., sum of relevance divided by the size of the brain region), and identify the most important features based on these two metrics.

In order to assign relevance to a corresponding brain area, one needs to perform a registration of the brains. We registered the brains to the 1 mm resolution 2009c version of the ICBMI152 reference brain, since this reference brain was used by Böhle et al,[19] which, to the best of our knowledge, is the only paper using LRP for AD diagnosis.

3 | RESULTS

3.1 | Classification results

The proposed 3D-CNN was trained for a total of 5 epochs with a batch size of 5, which was the maximum batch size our GPU could handle. The model achieved an accuracy of 90.57% for AD versus NC on the test dataset. The splitting of the dataset into train, test, and validation sets is shown in Table 3.

We further tried to increase the performance of the model by finding the optimal classification threshold, since there is a class imbalance towards the HC class. The optimal threshold was found via the precision-recall curve, as according to David et al[30] and Saito et al,[31] this is the preferred method in case of a moderate to large class imbalance. The identified best threshold via the precision-recall curve was 0.4241 with an F score of 0.8877. After changing the classification threshold of the model, the AD versus HC accuracy improved from 90.52% to 92.11%.

The recall score of our trained model was 0.96 for the HC class, which means that in 96% of the cases, the model correctly classified a healthy subject as HC. The recall score of the AD class was 0.86, which means that the model correctly identified the disease in 86% of the AD patients.

We can conclude based on the results that our proposed 3D-CNN achieved similar classification accuracy for the AD versus HC case than other state-of-the-art works, but achieved better results than the majority of related works that relied only on the MRI modality. We can see in Table 4 that works that achieved better

TABLE 3 The dataset distribution of the subjects and MRI scans with respect to various sets used for training the proposed model

	Subjects	HC	AD	MRI	HC	AD
All	969	462	507	3634	2033	1601
Train	833	393	440	3110	1718	1392
Val	67	34	33	258	146	112
Test	69	35	34	266	169	97

TABLE 4 Comparison of our results to state-of-the-art approaches from the model accuracy perspective

Paper	Modality	Model	AD versus HC
Mehmood et al[10]	MRI	CNN	98.73%
Payan et al[6]	MRI	SAE + 3D-CNN	95.39%
Suk et al[32]	MRI, PET	DBM	95.35%
Liu et al[8]	MRI	3D-CNN	91.09%
Böhle et al[19]	MRI	3D-CNN	90.57%
ADNet (our)	MRI	3D-CNN	92.11%

Abbreviations: DBM, deep Boltzmann machine; RNN, recurrent neural network; SAE, sparse autoencoder.

accuracy than our proposed method used either a combination of multiple modalities (Suk et al[32]), used a combination of multiple models (Payan et al[6]), or used transfer learning with a combination of tissue segmentation (Mehmood et al[10]).

3.2 | Results of the positive evidence visualization

First, we present the results of the positive evidence visualization in the form of heatmaps. We compare the proposed composite LRP approach with the gradient-based sensitivity analysis and the uniform LRP. To identify the best composite LRP rule configuration, the pixel-flipping evaluation method was used where we iteratively replaced voxels with noise starting with the most relevant voxels first. Since in the most relevant first pixel-flipping approach the faster and further the classification score curve decreases the better, we chose the composite LRP configurations with the lowest area under the curve (AUC) scores. The top three configurations with the lowest AUC scores consisted of the same propagation rules, except for the input layer:

- Layer: Conv3D 2–5: LRP-$\alpha\beta$ ($\alpha = 1, \beta = 0$)
- Layer: Dense 1: LRP-$\alpha\beta$ ($\alpha = 1, \beta = 0$)
- Layer: Dense 2–3: LRP-ϵ
- Layer: Dense 4 (output): LRP-0

For the input layer, we use the existing w^2 rule and the proposed $input \times w^2$ and $\sqrt{w}$ modifications. The positive evidence interpretations are shown in Figure 2.

The first observation we can make about the visualizations is that the sensitivity analysis exhibits a significant background noise. This can be explained by how the gradient-based analysis works. During the training process, the background consists of mostly zeros (or almost zeros) and looks the same for all the samples; therefore, the model does not consider it as a representative feature and ignores it. During the analysis, however, when the method is measuring how a value change in the background affects the output score, the model does not know how to interpret it, since it had never encountered a background value different than zero and the prediction will be highly disturbed.

The LRP approaches visually look almost identical except for the composite LRP with the $input \times w^2$ input rule. Although this variant shows relevance in the same brain areas, it shows less relevance overall and is generally sparser. This might be due to the fact that the input often contains values less than one, which then dampens the relevance after the multiplication. The advantage of this is that the interpretation will only show the most relevant regions by filtering out the least relevant ones, making the visualization more focused.

The interpretations also correspond to the literature, as all the identified brain regions are relevant AD biomarkers. Each technique shows relevance in the middle temporal gyrus (MTG), inferior temporal gyrus (ITG), ventricles, and the hippocampal area. The LRP methods also show significant relevance in the cerebellum region, which is responsible for movement and balance, and the optic nerves alongside the optic chiasm, which were proven to be AD biomarkers by Nishioka et al[33] and Armstrong.[34]

3.3 | Results of the negative evidence visualization

In case of the negative evidence visualization, we compare three LRP approaches: Uniform LRP, composite LRP, and modified top layer approach. For the uniform LRP, we chose the $\alpha\beta$-rule with $\alpha = 2$ and $\beta = 1$, as this is capable of showing the negative evidence alongside the positive evidence. For the composite LRP and modified top layer, we used the pixel-flipping approach to determine the best configurations, similarly as in case of the positive evidence visualization. The rule configuration with the lowest AUC for composite LRP:

- Layer: Conv3D 2–5: LRP-$\alpha\beta$ ($\alpha = 2, \beta = 1$)
- Layer: Dense 1: LRP-$\alpha\beta$ ($\alpha = 2, \beta = 1$)
- Layer: Dense 2–3: LRP-ϵ
- Layer: Dense 4 (output): LRP-0

The best configuration for the modified top layer approach:

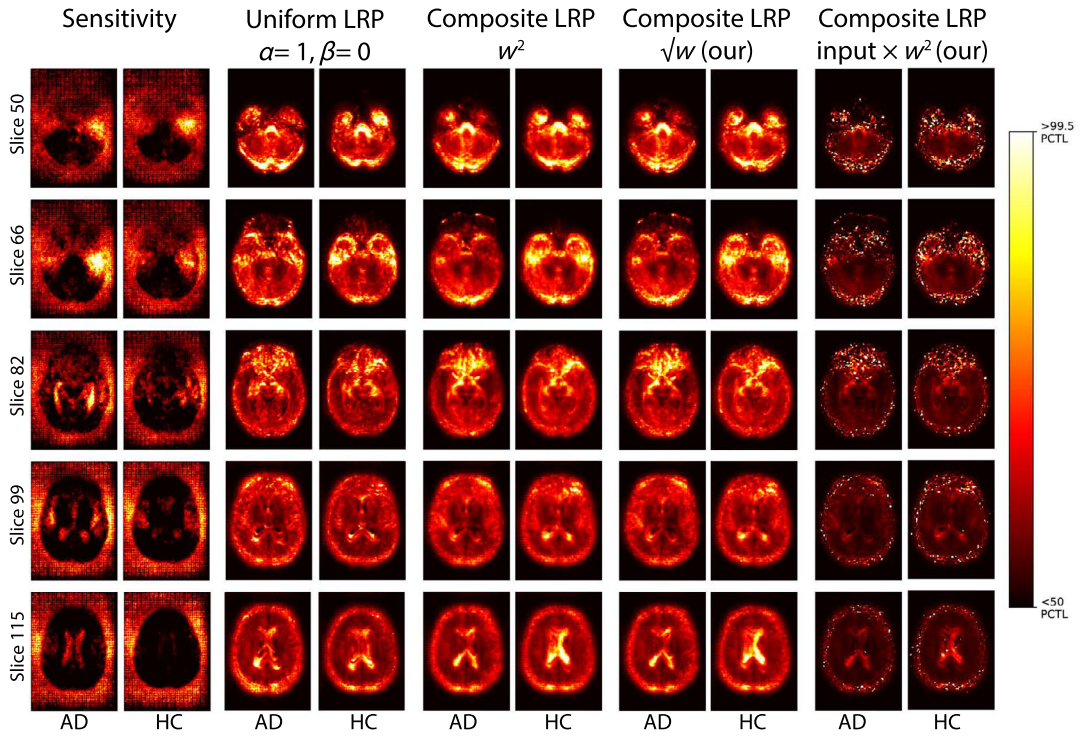

FIGURE 2 Comparison of the different interpretation methods for positive evidence visualization. All the analyzed input rules for composite LRP generate almost identical heatmaps, while the *input* $\times$ w^2 rule contains less relevance overall and is sparser, resulting in a more focused interpretation. The brain areas identified as relevant are all known AD biomarkers. Furthermore, the LRP approaches compared with sensitivity analysis show significant relevance in the optic nerves and optic chiasm, which are also proven AD biomarkers. The uniform LRP represents the LRP-$\alpha\beta$ rule with $\alpha = 1$ and $\beta = 0$, while the composite LRP represents the configuration described in subsection 3.2 with the existing w^2 and the proposed $\sqrt{w}$ and *input* $\times$ w^2 input rules. The heatmaps were obtained from 25 HC and 25 AD subjects with true positive classification results. Values between the 50th and 99.5th percentile are linearly color-coded, while values below and above the given percentiles are black, resp. white, as in the work of Böhle et al.[19]

- Layer: Conv3D 2–5: LRP-$\alpha\beta$ ($\alpha = 2, \beta = 1$)
- Layer: Dense 1–2: LRP-ϵ
- Layer: Dense 3–4: LRP-0

For the input layer, we used the bounded rule and our proposed *w*-log rule. The results of the negative evidence interpretations are shown in Figure 3.

We can observe from the results obtained via uniform LRP approach that the interpretations show negative evidence only in the case of the AD class. The only brain regions that are presented as a positive contribution for the HC class and as a negative contribution for the AD class are the middle and inferior temporal gyri.

In the case of the composite LRP approach, the interpretations exhibit negative evidence in both classes. The main difference between the two input rules is that while the bounded rule clearly differentiates between positive and negative regions, the proposed *w*-log rule considers every relevant region (with a few exceptions) equally important for both the AD and HC classes. The most noticeable difference between the two evidence types in the case of the *w*-log rule is seen in Slice 50, where the temporal lobe is presented as a positive contribution to the HC class.

If we compare the uniform LRP and composite LRP approaches, we can find some interesting patterns. The most obvious one is seen in Slice 50, where the right temporal lobe clearly supports the HC class (notice how the temporal lobe is shown is red in case of the HC class, and in blue in case of the AD class). This might indicate the nature of the disease, since damage to the left temporal lobe means the patient has problems with verbal semantic memory, while damage to the right side affects the visual memory (e.g., recall of faces) of the patient.

Another interesting observation we can make about the interpretations is that there is a disagreement between the uniform and bounded approach regarding the temporal lobe seen in Slice 66 of the AD class. While the uniform LRP clearly presents the temporal lobe as a region supporting the HC class, the bounded rule marks it as a region supporting the AD class. Although the *w*-log rule seems to support the decision of the uniform rule

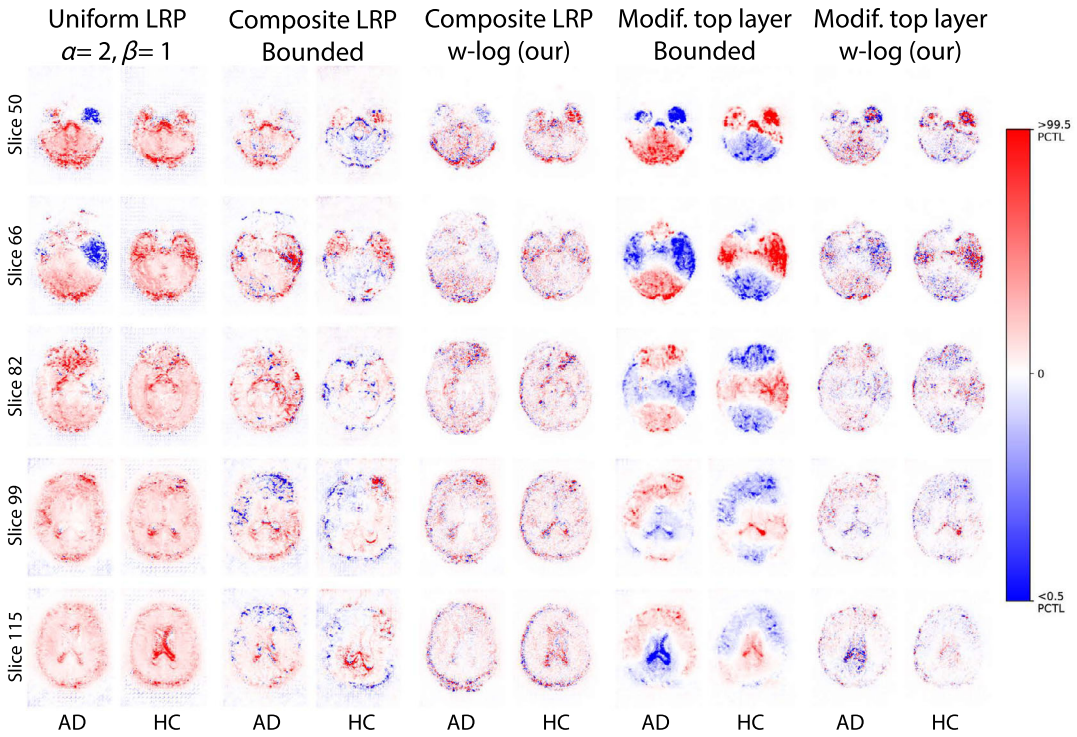

FIGURE 3 Comparison of the different interpretation methods for negative evidence visualization. The uniform LRP shows negative evidence only in the area of middle and inferior temporal gyri and only in the case of the AD class, while the composite LRP and modified top layer approaches exhibit much more negative evidence overall. Additionally, the bounded rule strictly differentiates between AD and HC regions, especially in the case of the modified top layer. The uniform LRP represents the LRP-$\alpha\beta$ rule with $\alpha = 2$ and $\beta = 1$, while the composite LRP represents the configurations described in subsection 3.3 with the input rules bounded and our proposed $w - log$ rule. The heatmaps were obtained from 25 HC and 25 AD subjects with true positive classification results. The colormap is centered around zero, where the red color corresponds to the positive contributions of the predicted class and the blue color to the negative contributions. Values above the 99.5th percentile and below the 0.5th percentile red, resp. blue

(i.e., the temporal lobe contains evidence supporting the HC class), it is not evident, and further investigation is required.

From the interpretations obtained via the modified top layer approach, it is clear which brain region corresponds to which class, as one class is the exact opposite of the other class from the relevance point of view. This is an improvement compared with the uniform or composite LRP approaches, where one class was not necessarily the exact opposite of the other class from the visualization point of view.

Another improvement the output layer modification has brought is the elimination of the background relevance in the heatmaps. Although very little relevance is still present in the background of the AD class when using the bounded rule, it is much more subtle than in the case of the composite approach.

If we compare the bounded and w-log input rules, we can see that they agree on which brain regions support the HC class and the AD class. According to the modified top layer approach, the temporal lobe, the lateral ventricles, and the whole middle section of the brain contain evidence that supports the HC class. The AD class, on the other hand, is mainly supported by the frontal

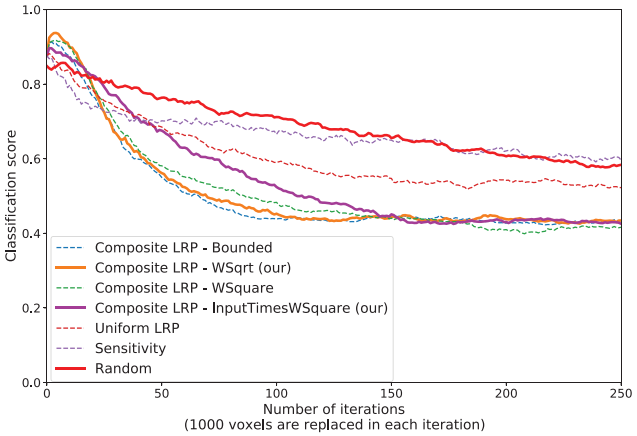

FIGURE 4 Comparison of different interpretation methods used for visualizing positive evidence via the most relevant pixel-flipping approach. The classification score represents the average score acquired from 25 HC and 25 AD subjects

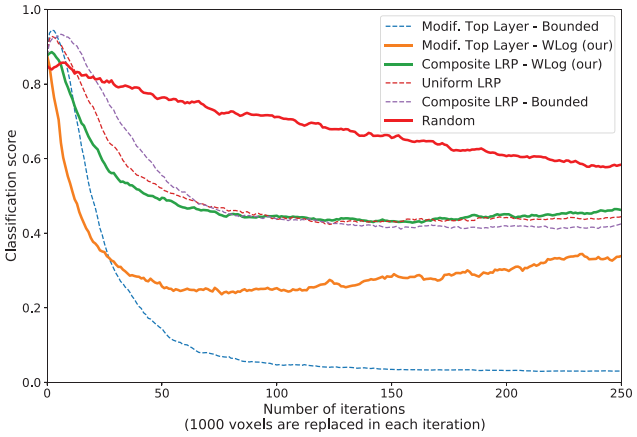

FIGURE 5 Comparison of different interpretation methods used for visualizing negative evidence via the most relevant pixel-flipping approach. The classification score represents the average score acquired from 25 HC and 25 AD subjects

lobe, which controls important cognitive skills in humans, such as memory, emotional expression, problem-solving, and language, the occipital lobe, which is the visual processing center, and the cerebellum, which is responsible for balance, coordination, and fine muscle control.

3.4 | Results of the pixel-flipping evaluation

In the quantitative pixel-flipping analysis of the proposed interpretation methods, we chose to replace 1000 voxels in each iteration, while the total number of iterations was set to 250. The voxels were replaced with random noise from a uniform distribution, which was bounded by the minimum and maximum voxel values of the analyzed volume.

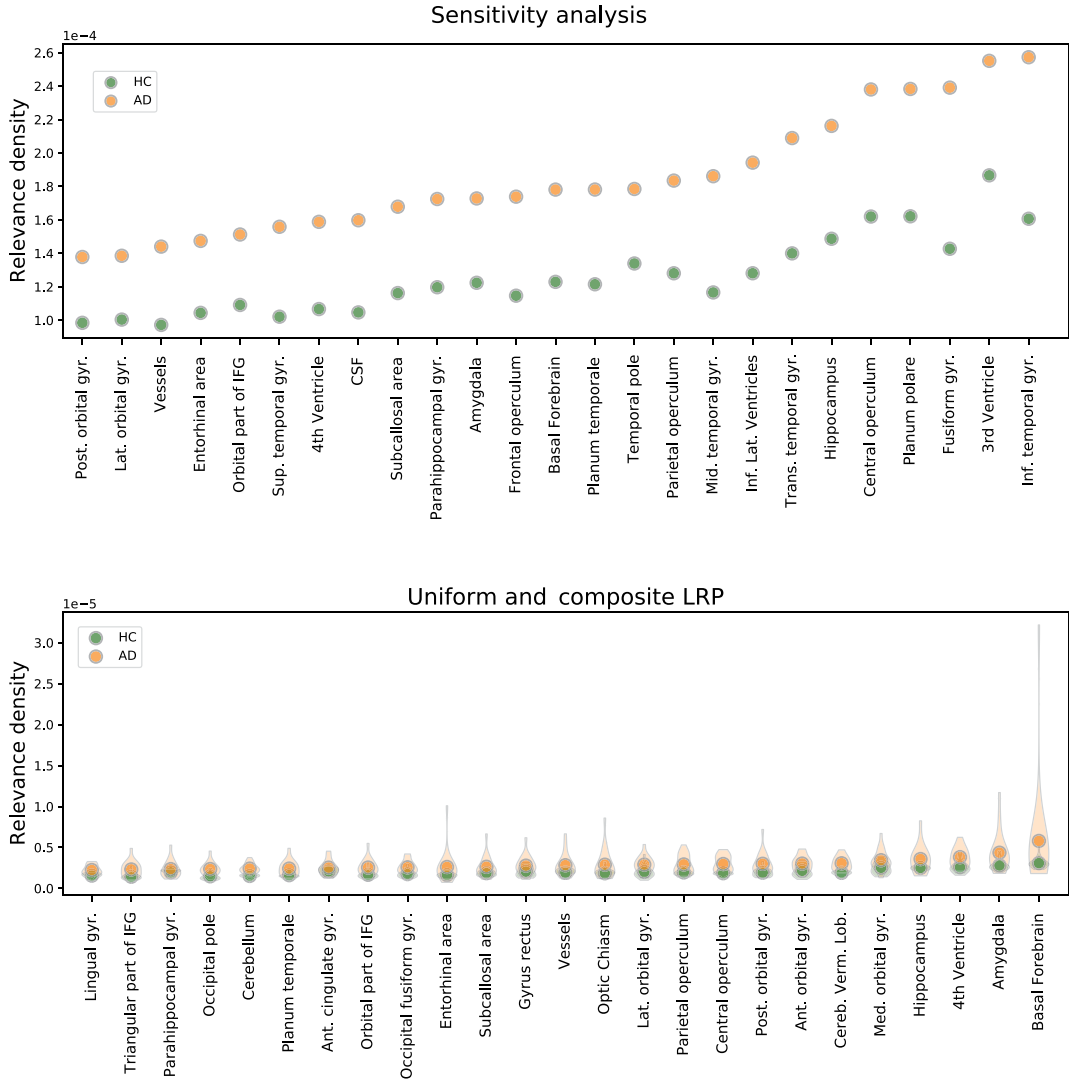

FIGURE 6 Top 25 most relevant brain areas for each of the respective methods used for visualizing positive evidence. The uniform and composite LRP approaches showed almost identical results, therefore we decided to evaluate them together. The results were obtained from 20 HC and 20 AD individuals. The visualizations were inspired by Böhle et al.[19]

Our results conclude that in the case of the interpretation methods, which visualize positive evidence, the composite LRP approach outperforms the uniform LRP or the gradient-based sensitivity analysis. Furthermore, the proposed $\sqrt{w}$ input rule achieved better results than the existing w^2-rule, which we based our modifications on. The $input \times w^2$-rule, on the other hand, did not show improvement compared with the w^2-rule, but still outperformed the uniform LRP and the sensitivity analysis. The results are shown in Figure 4.

The results also support the findings of Böhle et al,[19] who demonstrated that the relevance-based interpretation methods perform better than the gradient-based methods. The interpretation approaches proposed by us outperformed their uniform LRP approach, concluding that the composite LRP (with specific configuration) is a better interpretation method for visualizing positive evidence in neural networks focusing on AD classification based on MRI images.

In case of heatmaps containing negative evidence, there are two potential approaches on how to sort the voxels, since one might argue that the negatively relevant voxels are still as relevant as the positively relevant voxels, and a change in the negatively relevant regions might still highly affect the classification score. For this reason, one can sort the voxels either based on their raw or their absolute relevance score. We decided to sort the voxels based on their raw relevance score.

We can observe from the results shown in Figure 5 that the modified top layer approach achieved marginally better results than the composite or uniform LRP approaches. Although the lowest AUC score was

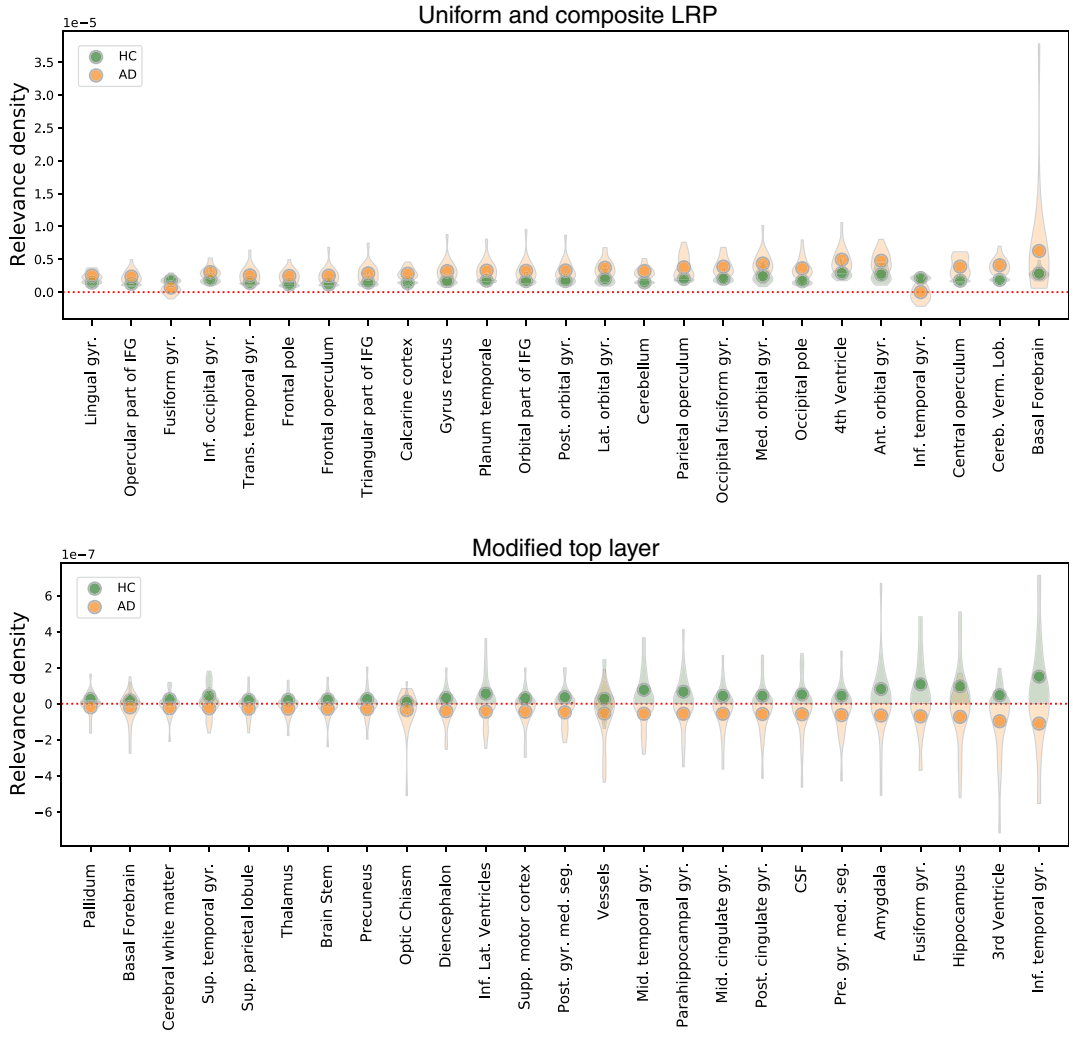

FIGURE 7 Top 25 most distinctive brain areas for each of the respective methods used for visualizing negative evidence. The modified top layer approach is better at distinguishing positive and negative contributions than the uniform and composite LRP approaches. Since the uniform and composite LRP approaches showed almost identical results, we decided to evaluate them together. The same is true for the modified top layer approach, where the results represent the average result obtained via the bounded and *w*-log input rules. The results were obtained from 20 HC and 20 AD individuals

achieved by the modified top layer method with the existing bounded input rule, our proposed *w*-log rule is better at identifying the most relevant features in the beginning but fails to keep up with the bounded rule later on.

We can also notice how the classification scores of the configurations containing the *w*-log rule go up after approx. 100 iterations. This might be caused by the fact that the *w*-log rule does not differentiate between positive and negative regions as clearly as the bounded rule, therefore often the most positive and the most negative contributions are in the same brain region. This suggests that specific regions might be equally important for both classes.

3.5 | Results of the Atlas-based evaluation

The results of the Atlas-based evaluation for interpretations methods used for visualizing positive evidence are shown in Figure 6. The LRP approaches were merged into a single figure, as the results were almost identical for both uniform and composite LRP methods.

This quantitative Atlas-based evaluation also confirms that the interpretation methods identified relevant brain regions, while the LRP approach considers the hippocampus and amygdala, which are perhaps the most notable early AD biomarkers, more important than the sensitivity analysis. This further indicates the advantages of the LRP methods in contrast to the gradient-based methods.

We would also like to point out that our results of the sensitivity analysis contradict the findings of Böhle et al,[19] where the sensitivity analysis did not differentiate between HC and AD class from the relevance density point of view. In our case, however, the two classes can be clearly differentiated from each other, as the AD class exhibits more relevance than the HC class for the respective brain areas.

The size-normalized relevance for the interpretation methods visualizing negative evidence is shown in Figure 7. The results confirm our findings from the heatmaps, according to which the uniform and composite LRP methods mainly exhibit negative evidence in the ITG. Moreover, the fact that the modified top layer approach clearly distinguishes between HC and AD brain areas is also evident in this figure.

4 | CONCLUSION

In this paper, a novel neural network approach was proposed to diagnose AD based on MRI scans alongside various interpretation methods to explain the decision of the classifier with the goal to make the computer-aided diagnosis via neural networks more viable by eliminating their "black-box" nature. The proposed 3D-CNN classifier achieved an accuracy of 92.11% for AD versus HC, which is comparable to state-of-the-art approaches and performs better than the majority of models that rely only on MRI scans and a single model.

The proposed LRP interpretation approaches can be divided into two categories: methods able to visualize only positive evidence, and methods able to visualize both positive and negative evidence. For the positive evidence, two propagation rules were proposed, namely $input \times w^2$ and $\sqrt{w}$, and also the application of composite LRP. The results were compared with sensitivity analysis and uniform LRP, and based on the quantitative evaluation, the composite LRP outperformed both these techniques, while the proposed $input \times w^2$ rule outperformed the existing w^2 rule, which the modification was based on.

As for the negative evidence, the *w*-log rule was proposed alongside the application of composite LRP and modified top layer approaches. Although the composite LRP approach had a hard time identifying negative regions, the modified top layer approach clearly distinguished between positive and negative brain areas, which further helps the radiologists to deliver the correct diagnosis.

The proposed model, therefore, can be used to aid the radiologists to identify AD based on MRI scans, while the proposed interpretation methods can be used to explain the decision of the model, which is necessary in the medical field. Additionally, the proposed interpretation methods are model-agnostic, meaning that they can be applied to arbitrary CNN model in the arbitrary domain and therefore have a wide range of applications, which helps the spread of computer-aided diagnosis by reducing the "black-box" nature of neural-networks.

ACKNOWLEDGMENTS

The authors would like to thank for financial contribution from the STU Grant scheme for Support of Young Researchers. Data collection and sharing for this project were funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and Department of Defense (DOD) ADNI (Award No: 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

Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu) and is publicly available for researchers.

ORCID

Tomáš Pohl https://orcid.org/0000-0002-4695-0285
Marek Jakab https://orcid.org/0000-0002-4329-6417
Wanda Benesova https://orcid.org/0000-0001-6929-9694

ENDNOTES

*adni.loni.usc.edu.

†https://github.com/tomaspohl/innvestigate.

REFERENCES

1. Blank RH. *Alzheimer's Disease and Other Dementias: An Introduction*. Springer Singapore; 2019:1-26.
2. Wattamwar PR, Mathuranath PS. An overview of biomarkers in Alzheimer's disease. *Ann Indian Acad Neurol*. 2010;13(Suppl 2):116-123.
3. Corina P, Miia K, Susanna T, et al. Hippocampus and entorhinal cortex in mild cognitive impairment and early AD. *Neurobiol Aging*. 2004;25:303-310.
4. Kishore DB. *Positron Emission Tomography: An Overview*. Springer India; 2015:1-6.
5. Johnson KA, Fox NC, Sperling RA, Klunk WE. Brain imaging in Alzheimer disease. *Cold Spring Harb Perspect Med*. 2012; 2(4):a006213.
6. Adrien P, Giovanni M. Predicting Alzheimer's disease: a neuroimaging study with 3D convolutional neural networks. CoRR; 2015. arXiv:1502.02506.
7. Manhua L, Danni C, Weiwu Y. Alzheimer's disease neuroimaging initiative. Classification of Alzheimer's disease by combination of convolutional and recurrent neural networks using FDG-PET images. *Front Neuroinf*. 2018;12:35.
8. Mingxia L, Jun Z, Ehsan A, Shen D. Landmark-based deep multi-instance learning for brain disease diagnosis. *Med Image Anal*. 2018;43:157-168.
9. Manhua L, Fan L, Hao Y, et al. A multi-model deep convolutional neural network for automatic hippocampus segmentation and classification in Alzheimer's disease. *Neuro Image*. 2020;208:116459.
10. Atif M, Yang S, Zhixi F, et al. A transfer learning approach for early diagnosis of Alzheimer's disease on MRI images. *Neuroscience*. 2021;460:43-52.
11. Taeho J, Kwangsik N, Saykin Andrew J. Deep learning in Alzheimer's disease: diagnostic classification and prognostic prediction using neuroimaging data. *Front Aging Neurosci*. 2019;11:220.
12. Ebrahimighahnavieh A, Luo S, Chiong R. Deep learning to detect Alzheimer's disease from neuroimaging: a systematic literature review. *Comput Methods Programs Biomed*. 2020;187: 105242.
13. Tanveer M, Richhariya B, Khan RU, et al. Machine learning techniques for the diagnosis of Alzheimer's disease: a review. *ACM Trans. Multimedia Comput Commun Appl*. 2020;16(1s):1-35.
14. Sebastian L, Stephan W, Alexander B, Grégoire M, Wojciech S, Klaus-Robert M. Unmasking clever Hans predictors and assessing what machines really learn. *Nat Commun*. 2019;10(1):1096.
15. Bryce G, Seth F. European Union regulations on algorithmic decision-making and a "right to explanation". *AI Mag*. 2017; 38(3):50-57.
16. Karen S, Andrea V, Andrew Z. Deep inside convolutional networks: visualising image classification models and saliency maps. arXiv:1312.6034.2013.
17. Tobias SJ, Alexey D, Thomas B, Martin R. Striving for simplicity: the all convolutional net; 2014.
18. Avanti S, Peyton G, Anshul K. Learning important features through propagating activation differences. In: Doina P, Whye TY, eds. *Proceedings of the 34th International Conference on Machine Learning*, *Proceedings of Machine Learning Research*, vol. 70. International Convention Centre; 2017:3145–3153.
19. Moritz B, Fabian E, Martin W, Kerstin R. Layer-wise relevance propagation for explaining deep neural network decisions in MRI-based Alzheimer's disease classification. *Front Aging Neurosci*. 2019;11:194.
20. Zeiler MD, Rob F. Visualizing and understanding convolutional networks. In: David F, Tomas P, Bernt S, Tinne T, eds. *Computer Vision–ECCV 2014*. Springer International Publishing; 2014:818-833.
21. Johannes R, Fabian E, Martin W, John-Dylan H, Kerstin R. Visualizing convolutional networks for MRI-based diagnosis of Alzheimer's disease. *Lect Notes Comput Sci*. 2018;11038:24-31.
22. Sebastian B, Alexander B, Grégoire M, Frederick K, Klaus-Robert M, Wojciech S. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS One*. 2015;10(7):e0130140.
23. Grégoire M, Alexander B, Sebastian L, Wojciech S, Klaus-Robert M. *Layer-Wise Relevance Propagation: An Overview*. Springer International Publishing; 2019:193-209.
24. Marinescu RV, Oxtoby NP, Young AL, et al. TADPOLE challenge: prediction of longitudinal evolution in Alzheimer's disease; 2018.

25. Maximilian A, Sebastian L, Philipp S, et al. iNNvestigate neural networks! *J Mach Learn Res*. 2019;20(93):1-8.
26. Grégoire M, Sebastian L, Alexander B, Wojciech S, Klaus-Robert M. Explaining nonlinear classification decisions with deep Taylor decomposition. *Pattern Recognit*. 2017;65:211-222.
27. Jacob K, Malte E, Grégoire M, Wojciech S, Klaus-Robert M. From clustering to cluster explanations via neural networks; 2019. ArXiv:abs/1906.07633.
28. Samek W, Binder A, Montavon G, Lapuschkin S, Müller K. Evaluating the visualization of what a deep neural network has learned. *IEEE Trans Neural Netw Learn Syst*. 2017;28(11):2660-2673.
29. Alexander B, Grégoire M, Sebastian L, Klaus-Robert M, Wojciech S. Layer-wise relevance propagation for neural networks with local renormalization layers. *Lect Notes Comput Sci*. 2016;9887 LNCS:63-71.
30. Jesse D, Mark G. The relationship between precision-recall and ROC curves. In: *ICML'06: 233–240Association for Computing Machinery*; 2006.
31. Takaya S, Marc R. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PloS One*. 2015;10(3):e0118432.
32. Heung-Il S, Seong-Whan L, Dinggang S. Initiative Alzheimer's disease neuroimaging. Hierarchical feature representation and multimodal fusion with deep learning for AD/MCI diagnosis. *Neuro Image*. 2014;101:569-582.
33. Christopher N, Christina P, Shu-Wei S. Diffusion tensor imaging reveals visual pathway damage in patients with mild cognitive impairment and Alzheimer's disease. *J Alzheimer's Dis*. 2015;45(1):97-107.
34. Armstrong RA. Alzheimer's disease and the eye. *J Optom*. 2009;2(3):103-111.

How to cite this article: Pohl T, Jakab M, Benesova W. Interpretability of deep neural networks used for the diagnosis of Alzheimer's disease. *Int J Imaging Syst Technol*. 2021;1-14. doi: 10.1002/ima.22657