Improving 3D convolutional neural network comprehensibility via interactive visualization of relevance maps: evaluation in Alzheimer’s disease

Martin Dyrba1*, Moritz Hanzig1,2, Slawek Altenstein3,4, Sebastian Bader2, Tommaso Ballarini5, Frederic Broserson5,6, Katharina Buerger7,8, Daniel Cantré9, Peter Dechent10, Laura Dobisch11, Emrah Düzel11,12, Michael Ewers7,8, Klaus Fliessbach5,6, Wenzel Glanz11,13, John-Dylan Haynes13, Michael T. Heneka5,6, Daniel Janowitz8, Deniz B. Keles14, Ingo Kilimann1,15, Christoph Laske16,17,18, Franziska Maier19, Coraline D. Metzger11,12,20, Matthias H. Munk16,18,21, Robert Perneckzy7,22,23, Oliver Peters3,14, Lukas Preis3,14, Josef Priller3,4,25, Boris Rauchmann22, Nina Roy5, Klaus Scheffler26, Annika Spotteke3,4, Eike J. Spruth3,4, Marc-André Weber9, Birgit Ertl-Wagner3,4, Michael Wagner5,6, Jens Wiltfang27,28,33, Frank Jessen5,19,34 and Stefan J. Teipel1,15 for the ADNI, AIBL, DELCODE study groups

Abstract

Background: Although convolutional neural networks (CNNs) achieve high diagnostic accuracy for detecting Alzheimer’s disease (AD) dementia based on magnetic resonance imaging (MRI) scans, they are not yet applied in clinical routine. One important reason for this is a lack of model comprehensibility. Recently developed visualization methods for deriving CNN relevance maps may help to fill this gap as they allow the visualization of key input image features that drive the decision of the model. We investigated whether models with higher accuracy also rely more on discriminative brain regions predefined by prior knowledge.

Methods: We trained a CNN for the detection of AD in N = 663 T1-weighted MRI scans of patients with dementia and amnestic mild cognitive impairment (MCI) and verified the accuracy of the models via cross-validation and in three independent samples including in total N = 1655 cases. We evaluated the association of relevance scores and hippocampus volume to validate the clinical utility of this approach. To improve model comprehensibility, we implemented an interactive visualization of 3D CNN relevance maps, thereby allowing intuitive model inspection.

Results: Across the three independent datasets, group separation showed high accuracy for AD dementia versus controls (AUC ≥ 0.91) and moderate accuracy for amnestic MCI versus controls (AUC ≈ 0.74). Relevance maps indicated that hippocampal atrophy was considered the most informative factor for AD detection, with additional contributions from atrophy in other cortical and subcortical regions. Relevance scores within the hippocampus were highly correlated with hippocampal volumes (Pearson’s r ≈ −0.86, p < 0.001).

*Correspondence: martin.dyrba@dzne.de
1 German Center for Neurodegenerative Diseases (DZNE), Rostock, Germany
Full list of author information is available at the end of the article

© The Author(s) 2021. Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated in a credit line to the data.
Conclusion: The relevance maps highlighted atrophy in regions that we had hypothesized a priori. This strengthens the comprehensibility of the CNN models, which were trained in a purely data-driven manner based on the scans and diagnosis labels. The high hippocampus relevance scores as well as the high performance achieved in independent samples support the validity of the CNN models in the detection of AD-related MRI abnormalities. The presented data-driven and hypothesis-free CNN modeling approach might provide a useful tool to automatically derive discriminative features for complex diagnostic tasks where clear clinical criteria are still missing, for instance for the differential diagnosis between various types of dementia.

Keywords: Alzheimer’s disease, Deep learning, Convolutional neural network, MRI, Layer-wise relevance propagation

Introduction

Alzheimer’s disease (AD) is characterized by widespread neuronal degeneration, which manifests macroscopically as cortical atrophy that can be detected in vivo using structural magnetic resonance imaging (MRI) scans. Particularly at earlier stages of AD, atrophy patterns are relatively regionally specific, with volume loss in the medial temporal lobe and particularly the hippocampus. Therefore, hippocampus volume is currently the best-established MRI marker for diagnosing Alzheimer’s disease at the dementia stage as well as at its prodromal stage amnestic mild cognitive impairment (MCI) [1, 2]. Automated detection of subtle brain changes in early stages of Alzheimer’s disease could improve diagnostic confidence and early access to intervention [1, 3].

Convolutional neural networks (CNNs) provide a powerful method for image recognition. Various studies have evaluated the performance of CNNs for the detection of Alzheimer’s disease in MR images with promising results regarding both separation of diagnostic groups and the prediction of conversion from MCI to manifest dementia. Despite the high accuracy levels achieved by CNN models, a major drawback is their algorithmic complexity, which renders them black-box systems. The poor intuitive comprehensibility of CNNs is one of the major obstacles which hinder the clinical application.

Novel methods for deriving relevance maps from CNN models [4, 5] may help to overcome the black-box problem. In general, relevance or saliency maps indicate the amount of information or contribution of a single input feature on the probability of a particular output class. Previous methodological approaches like gradient-weighted class activation mapping (Grad-CAM) [6], occlusion sensitivity analyses [7, 8], and local interpretable model-agnostic explanations (LIME) [9] had the limitation that deriving the relevance or saliency maps provided only group-average estimates, required long runtime [10], or provided only low spatial resolution [11, 12]. In contrast, more recent methods such as guided backpropagation [13] or layer-wise relevance propagation (LRP) [4, 5] use back-tracing of neural activation through the network paths to obtain high-resolution relevance maps.

Recently, three studies compared LRP with other CNN visualization methods for the detection of Alzheimer’s disease in T1-weighted MRI scans [11, 12, 14]. The derived relevance maps showed the strongest contribution of medial and lateral temporal lobe atrophy, which matched the a priori expected brain regions of high diagnostic relevance [15, 16]. These preliminary findings provided the first evidence that CNN models and LRP visualization could yield reasonable relevance maps for individual people. We investigated whether this approach could be used as the basis for neuroradiological assistance systems to support the examination and diagnostic evaluation of MRI scans. Furthermore, we wanted to develop a data-driven and hypothesis-free CNN modeling approach that is capable of automatically deriving discriminative features and, therefore, might support complex diagnostic tasks where clear clinical criteria are still missing such as the differential diagnosis of various types of dementia.

In the current study, our aims were threefold: First, we trained robust CNN models that achieved a high diagnostic accuracy in three independent validation samples. Second, we developed a visualization software to interactively derive and inspect diagnostic relevance maps from CNN models for individual patients. Here, we expected high relevance to be shown in brain regions with strong disease-related atrophy, primarily in the medial temporal lobe. Third, we evaluated the validity of relevance maps in terms of correlation of hippocampus relevance scores and hippocampus volume, which is the best-established MRI marker for Alzheimer’s disease [15, 16]. We expected a high consistency of both measures, which would strengthen the overall comprehensibility of the CNN models.

State of the art

Neural network models to detect Alzheimer’s disease

An overview of neuroimaging studies which applied neural networks in the context of AD is provided in
Table 1. We focused on the aspects whether the studies used independent validation samples to assess the generalizability of their models and whether they evaluated which image features contributed to the models’ decision. Studies reported very high classification performances to differentiate AD dementia patients and cognitively healthy participants, typically with accuracies around 90% (Table 1). For the separation of MCI and controls, accuracies were substantially lower ranging between 75 and 85%. However, there is a high variation of the accuracy levels depending on various factors such as (i) differences in diagnostic criteria across samples, (ii) included data types, (iii) differences in image preprocessing procedures, and (iv) differences between machine learning methods [27].

CNN performance estimation and model robustness are still open challenges. Wen and colleagues [27] actually showed only a minor effect of the particular CNN model parameterization or network layer configuration on the final accuracy, which means that the fully trained CNN models achieved almost identical performance. Different CNN approaches exist for MRI data [27] based on (i) 2D convolutions for single slices, often reusing pre-trained models for general image detection, such as AlexNet [29] and VGG [30]; (ii) so-called 2.5D approaches running 2D convolutions on each of the three slice orientations, which are then combined at higher layers of the network; and (iii) 3D convolutions, which are at least theoretically superior in detecting texture and shape features in any direction of the 3D volume. Although final accuracy is almost comparable between all three approaches for detecting MCI and AD [27], the 3D models require substantially more parameters to be estimated during training. For instance, a single 2D convolutional kernel has $3 	imes 3 = 9$ parameters whereas the 3D version requires $3 	imes 3 	imes 3 = 27$ parameters. Here, relevance maps and related methods enable the assessment of learnt CNN models with respect to overfitting to clinically irrelevant brain regions and the detection of potential biases present in the training samples, which cannot be directly identified just from the model accuracy.

Approaches to assess model comprehensibility
In the literature, the most often applied methods to assess model comprehensibility and sensitivity were (i) the visualization of model weights, (ii) occlusion sensitivity analysis, and (iii) more advanced CNN methods such as guided backpropagation or LRP (Table 1). Notably, studies using approaches i and ii showed visualizations characterizing the whole sample or group averages. In contrast, studies applying iii also presented relevance maps for single participants [11, 14]. Böhle and colleagues [14] pioneered the application of LRP in neuroimaging and reported a high sensitivity of this method to actual regional atrophy. Eitel and colleagues [12] assessed the stability and reproducibility of CNN performance results and LRP relevance maps. After training ten individual models based on the same training dataset, they reported the highest consistency and lowest deviation of relevance maps for LRP and guided backpropagation among five different methods [12]. Recently, we compared various methods for relevance and saliency attribution [11]. Visually, all tested methods provided similar relevance maps except for Grad-CAM, which provided much lower spatial resolution, and, hence, lost a high amount of regional specificity. For the other methods, the main difference was the amount of “negative” relevance which indicates evidence against a particular diagnostic class. Notably, [12, 14] did not include patients in the prodromal stage of MCI and [11] focused on a limited range of coronal slices covering the temporal lobe. All three studies did not validate their results in independent samples.

Materials and methods
Study samples
Data for training the CNN models were obtained from the Alzheimer’s Disease Neuroimaging Initiative (ADNI) database (https://adni.loni.usc.edu). The ADNI was launched in 2003 by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, the Food and Drug Administration, private pharmaceutical companies, and non-profit organizations, with the primary goal of testing whether neuroimaging, neuropsychological, and other biological measurements can be used as reliable in vivo markers of Alzheimer’s disease pathogenesis. A complete description of ADNI, up-to-date information, and a summary of diagnostic criteria are available at https://www.adni-info.org. We selected a sample of $N = 663$ participants from the ADNI-GO and ADNI-2 phases, based on the availability of concurrent T1-weighted MRI and amyloid AV45-PET scans. Notably, we used only one (i.e., the first) available scan from each ADNI participant in our analyses. The sample characteristics are shown in Table 2. We included 254 cognitively normal controls, 220 patients with (late) amnestic mild cognitive impairment (MCI), and 189 patients with Alzheimer’s dementia (AD). Amyloid-beta status of the participants was determined by the UC Berkeley [31] based on the AV45-PET standardized uptake value ratio (SUVR) cutoff 1.11.

For validation of the diagnostic accuracy of the CNN models, we obtained MRI scans from three independent cohorts. The sample characteristics and demographic information are summarized in Table 2. The


| Study (chronologic order) | Data type | Sample | Algorithm | Performance | Addressed model comprehensibility |
|---------------------------|-----------|--------|-----------|-------------|----------------------------------|
|                           |           | AD     | MCI       | Groups      | Accuracy | Balanced accuracy | AUC |                                    |
|                           |           |        | c/nc      | AD/CN       | MCI/CN   | MCI/MInc            |     |                                    |
|                           |           |        | CN        | AD/CN       | MCI/CN   | MCI/MInc            |     |                                    |
|                           |           |        |            |             |         |                   |     |                                    |
|                            |           |        |            |             |         |                   |     |                                    |
| Suk et al. [17]           | MRI GM and FDG-PET | 93     | 76/128    | 101         | RBM on class discriminative patches selected by statistical significance tests | AD/CN 95.4%  | 94.9% | MCI/CN 83.7%  | 80.6% | MCI/MInc 74.6%  | 71.6% | 0.988  | Visualization of selected features (image patches) and RBM model weights projected on MRI scan |
| Li et al. [18]            | MRI and FDG-PET   | 51     | 43/56     | 52          | RBM for feature learning, SVM for classification | AD/CN 91.4%  | 94.9% | MCI/CN 77.4%  | 71.6% | MCI/MInc 57.4%  | 51.6% | 0.881  | No |
| Ortiz et al. [19]         | MRI GM and FDG-PET | 70     | 39/64     | 68          | RBM for feature learning, SVM for classification | AD/CN 90%   | 88.2% | MCI/CN 83%    | 81%   | MCI/MInc 78%    | 75%   | 0.957  | No |
| Aderghal et al. [20]      | MRI and DTI       | 188    | 339       | 228         | CNN for hippocampus region of interest only | AD/CN 92.5%  | 92.5% | MCI/CN 80.0%  | 82.9% | MCI/MInc 78.9%  | 75%   | 0.957  | No |
| Liu et al. [21]           | FDG-PET         | 93     | 146       | 100         | CNN and RNN | AD/CN 91.2%  | 91%  | MCI/CN 78.9%  | 76%  | MCI/MInc 78.9%  | 75%  | 0.953  | Visualization of most contributing brain areas obtained from occlusion sensitivity analysis |
| Liu et al. [22]           | MRI             | 199    | –         | 229         | CNN on landmarks selected by statistical significance tests | AD/CN 90.6%  | 92.5% | MCI/CN 80.0%  | 82.9% | MCI/MInc 78.9%  | 75%  | 0.957  | Visualization of top 50 anatomical landmarks used as input for the CNN |
| Lin et al. [23]           | MRI             | 188    | 169/193   | 229         | CNN        | AD/CN 88.8%  | 88.8% | MCI/CN 79.9%  | 79%  | MCI/MInc 79.9%  | 75%  | 0.957  | No |
| Böhle et al. [14]         | MRI             | 211    | –         | 169         | CNN        | AD/CN 88.0%  | 88.0% | MCI/CN 79.9%  | 79%  | MCI/MInc 79.9%  | 75%  | 0.957  | Visualization of LRP relevance and guided backpropagation maps, comparison of LRP relevance scores by group and brain region |
| Li et al. [24]            | MRI Training 192 | –      | –         | –           | CNN for hippocampus only | AD/CN 92.9%  | 92.9% | MCI/CN 80.9%  | 82%  | MCI/MInc 78.9%  | 75%  | 0.957  | Visualization of most contributing hippocampus areas obtained from CNN class activation mapping |
| Dybra et al. [11]         | MRI Training 159 | –      | –         | –           | CNN for coronal slices covering hippocampus | AD/CN 90.3%  | 90.3% | MCI/CN 80.9%  | 82%  | MCI/MInc 78.9%  | 75%  | 0.957  | Visualization of LRP and other methods' relevance maps and comparison by diagnostic group |
| Lian et al. [25]          | MRI Training 199 | –      | –         | –           | CNN        | AD/CN 87.0%  | 87.0% | MCI/CN 76.6%  | 76%  | MCI/MInc 74.6%  | 74%  | 0.951  | Visualization of most contributing image areas obtained from CNN class activation mapping |
| Qiu et al. [26]           | MRI Training 188 | –      | –         | –           | FCN        | AD/CN 87.0%  | 87.0% | MCI/CN 76.6%  | 76%  | MCI/MInc 74.6%  | 74%  | 0.951  | Visualization of most contributing brain areas obtained from occlusion sensitivity analysis |
| Study (chronologic order) | Data type | Sample | Algorithm | Performance | Addressed model comprehensibility |
|---------------------------|-----------|--------|-----------|-------------|----------------------------------|
|                           |           |        | AD        | MCI c/nc     | CN                              |
|                           |           |        | Training 336 | 295/298 | CNN |
| Wen et al. [27]           | MRI       | Test1 76 | 20/13     | 429      | AD/CN1, MCIc/MCInc1, AD/CN2 |
|                           |           | Test2 78 | –         | 76       | 86% |
|                           |           |        | Test 76   | 330      | 50% |
|                           |           |        | Test 78   | –        | 70% |
| Thibeau-Sutre et al. [8]  | MRI       | Training 336 | – | 330 | CNN |
|                           |           | Test76 | –         | 429      | AD/CN |
|                           |           |        | Test 76   | 66       | 90% |
| Jo et al. [28]            | Tau-PET   | 66     | 66        | CNN      | AD/CN |

Empty cells in the performance columns indicate that the respective values were not reported.

AD = Alzheimer's dementia, MCI = mild cognitive impairment, MCIc = MCI converted to dementia, MCInc = non-converter/stable MCI, CN = cognitively normal controls, DTI = diffusion tensor imaging, FCN = fully connected network, RBM = restricted Boltzmann machine, RNN = recurrent neural network, CNN = convolutional neural network, MRI = T1-weighted magnetic resonance imaging, GM = gray matter volume, FDG-PET = glucose metabolism derived from fluorodeoxyglucose positron emission tomography.
Table 2  Summary of sample characteristics

| Sample                | CN   | MCI  | AD   |
|-----------------------|------|------|------|
| ADNI-GO/2 (training)  | N = 663 |
| Sample size (female)  | 254 (130) | 220 (93) | 189 (80) |
| Age (SD)              | 75.4 (6.6) | 74.1 (8.1) | 75.0 (8.0) |
| Education (SD)        | 16.4 (2.7) | 16.2 (2.8) | 15.9 (2.7) |
| MMSE (SD)             | 29.1 (1.2) | 27.6 (1.9) | 22.6 (3.2) |
| RAVLT Delayed recall (SD) | 7.6 (4.1) | 3.2 (3.7) | 0.8 (1.9) |
| WMS-LM Delayed recall (SD) | 13.9 (3.7) | 5.1 (3.8) | 1.5 (2.1) |
| Hippocampus volume (SD) | 6235 (756) | 5619 (963) | 4834 (930) |
| Amyloid status (neg/pos) | 177/77 | 79/141 | 28/161 |
| MRI field strength (1.5T/3T) | 71/183 | 49/171 | 35/154 |
| AIBL (validation) N = 606 |
| Sample size (female)  | 326 (211) | 187 (85) | 62 (27) |
| Age (SD)              | 70.0 (7.5) | 72.2 (7.5) | 74.8 (7.7) |
| Education (SD)        | 16.6 (2.2) | 16.6 (2.5) | 16.5 (2.4) |
| MMSE (SD)             | 29.1 (1.1) | 27.8 (2.0) | 23.1 (3.3) |
| RAVLT Delayed recall (SD) | 8.3 (4.4) | 4.7 (4.7) | 0.3 (0.9) |
| WMS-LM Delayed recall (SD) | 13.0 (3.5) | 7.2 (3.9) | 2.0 (2.8) |
| Hippocampus volume (SD) | 6583 (649) | 6112 (902) | 4839 (978) |
| Amyloid status (neg/pos) | 75/59 | 19/27 | 3/17 |
| MRI field strength (1.5T/3T) | 0/526 | 0/187 | 0/62 |
| DELCODE (validation) N = 575 |
| Sample size (female)  | 448 (260) | 96 (46) | 62 (36) |
| Age (SD)              | 72.4 (6.2) | 74.3 (6.9) | 73.2 (7.3) |
| Education (SD)        | 28.7 (1.2) | 27.0 (2.2) | 21.2 (5.3) |
| MMSE (SD)             | 11.2 (4.3) | 4.9 (4.0) | 1.0 (1.9) |
| RAVLT Delayed recall (SD) | 6362 (704) | 5712 (1028) | 4940 (1055) |
| Hippocampus volume (SD) | 316/101 | 34/54 | 6/53 |
| Amyloid status (neg/pos) | 55/393 | 7/89 | 2/60 |
| MRI field strength (1.5T/3T) | 5/104 |
| DELCODE (validation) N = 474 |
| Sample size (female)  | 215 (124) | 155 (72) | 104 (61) |
| Age (SD)              | 69.5 (5.5) | 73.0 (5.7) | 75.2 (6.2) |
| Education (SD)        | 14.7 (2.7) | 14.0 (3.1) | 12.9 (3.1) |
| MMSE (SD)             | 29.5 (0.8) | 27.8 (2.0) | 23.1 (3.2) |
| RAVLT Delayed recall (SD) | 14.3 (3.6) | 7.4 (5.2) | 1.8 (2.8) |
| Hippocampus volume (SD) | 6543 (679) | 5665 (950) | 4610 (944) |
| Amyloid status (neg/pos) | 58/28 | 30/57 | 5/49 |
| MRI field strength (1.5T/3T) | 0/215 | 0/155 | 0/104 |

Numbers indicate mean and standard deviation (SD) if not indicated otherwise.

Years of education were not available for the AIBL dataset. RAVLT Delayed recall scores were not available for the AIBL and DELCODE samples.

CN cognitively normal controls, MCI amnestic mild cognitive impairment, AD Alzheimer’s dementia, SD standard deviation, MMSE Mini Mental State Examination, RAVLT Rey Auditory Verbal Learning Test, WMS-LM Wechsler Memory Scale Logical Memory Test, MRI magnetic resonance imaging.

first dataset was compiled from \( N = 575 \) participants of the recent ADNI-3 phase. The second dataset included MR images from \( N = 606 \) participants of the Australian Imaging, Biomarker & Lifestyle Flagship Study of Ageing (AIBL) (https://aibl.csiro.au), provided via the ADNI system. A summary of the diagnostic criteria and additional information is available at https://aibl.csiro.au/about. For AIBL, we additionally obtained amyloid PET scans which were available for 564 participants (93%). The PET scans were processed using the Centiloid SPM pipeline and converted to Centiloid values as recommended for the different amyloid PET tracers [32–34]. Amyloid-beta status of the participants was determined using the cutoff 24.1 CL [33]. As a third sample, we included data from \( N = 474 \) participants of the German Center for Neurodegenerative Diseases (DZNE) multicenter observational study on Longitudinal Cognitive Impairment and Dementia (DELCODE) [35]. Comprehensive information on the diagnostic criteria and study design are provided in [35]. For the DELCODE sample, cerebrospinal fluid (CSF) biomarkers were available for a subsample of 227 participants (48%). Amyloid-beta status was determined using the \( A \beta/42/A \beta/40 \) ratio with a cutoff 0.09 [35].

Image preparation and processing

All MRI scans were preprocessed using the Computational Anatomy Toolbox (CAT12, v9.6/r7487) [36] for Statistical Parametric Mapping 12 (SPM12, v12.6/r1450, Wellcome Centre for Human Neuroimaging, London, UK). Images were segmented into gray and white matter, spatially normalized to the default CAT12 brain template in Montreal Neurological Institute (MNI) reference space using the DARTEL algorithm, resliced to an isotropic voxel size of 1.5 mm, and modulated to adjust for expansion and shrinkage of the tissue. Initially and after all processing steps, all scans were visually inspected to check for image quality. In all scans, effects of the covariates age, sex, total intracranial volume (TIV), and scanner magnetic field strength (FS) were reduced using linear regression. This step was performed, as these factors are known to affect the voxel intensities or regional brain volume [37,38]. For each voxel \( v_{xi} \), linear models were fitted on the healthy controls:

\[
v_{xi} = \beta_{i0} + \beta_{i1}age + \beta_{i2}sex + \beta_{i3}TIV + \beta_{i4}FS + \epsilon_{i}
\]  

with \( i \) being the voxel index, \( j \) being the healthy participant index, \( \beta_i \) being the respective model coefficients (for each voxel), and \( \epsilon_i \) being the error term or residual. Subsequently, the predicted voxel intensities were subtracted from all participants’ gray matter maps to obtain the residual images.
Notably, we performed the estimation process (1) only for the healthy ADNI-GO/2 participants. Then, (2) was applied to all other participants and the validation samples. This method was applied as brain volume, specifically in the temporal lobe and hippocampus, is substantially decreasing/shrinking in old age independently of the disease process [37, 38], and we expected this approach to increase accuracy. As sensitivity analysis, we also repeated CNN training on the raw gray matter volume maps for comparison. Patients with MCI and AD were combined into one disease-positive group. On the one hand, this was done as we observed a low sensitivity of machine learning models for MCI when trained only on AD cases, due to the much larger and more heterogeneous patterns of atrophy in AD than in MCI, where atrophy is specifically present in medial temporal and parietal regions [39]. On the other hand, combining both groups substantially increased the training sample, which was required to reduce the overfitting of the CNN models.

**CNN model structure and training**

The CNN layer structure was adapted from [14, 27], which was inspired by the prominent 2D image detection networks AlexNet [29] and VGG [30]. The model was implemented in Python 3.7 with Keras 2.2.4 and Tensorflow 1.15. The layout is shown in Fig. 1. The residualized/raw 3D images with a resolution of 100 × 100 × 120 voxels were fed as input into the neural network and processed by three consecutive convolution blocks including 3D convolutions (5 filters of 3 × 3 × 3 kernel size) with rectified linear activation function (ReLU), maximum pooling (2 × 2 × 2 voxel patches), and batch normalization layers (Fig. 1). Then, three dropout (10%) and fully connected layers with ReLU activation followed, each consisting of 64, 32, and 2 neurons, respectively. The weights of last two layers were regularized with the L2 norm penalty. The last layer had the softmax activation function that rescaled the class activation values to likelihood scores. The network required approximately 700,000 parameters to be estimated.

The whole CNN pipeline was evaluated by stratified tenfold cross-validation, partitioning the ADNI-GO/2 sample into approximately 600 training and 60 test images with almost equal distribution of CN, MCI, and AD cases. Additionally, data augmentation was used. All images included in the respective training subsamples were flipped along the coronal (L/R) axis and also translated by ±10 voxels in each direction (x/y/z), yielding fourteen times increased number of samples per epoch of approximately 8350 images. The CNN model was then trained with the ADAM optimizer, applying the categorical cross-entropy loss function, the learning rate of 0.0001, and a batch size of 20. As the training group sizes were imbalanced, we set class weights of 1.31 for controls and 0.81 for MCI/AD in order to circumvent biased predictions. The weights were determined using the formula 0.5n/ni as recommended in the TensorFlow tutorial [40]. To select the optimal models during training, we set the number of epochs to ten and saved the model state (epoch) which performed best on the test partition. On a Windows 10 computer with Intel Core i5-9600 hexa-core CPU, 64 GB working memory, and NVIDIA GeForce GTX 1650 CUDA GPU, training took approximately 35 min per fold and 12 h in total. All ten models were saved to disk for further inspection and validation. As control

![Data flow chart and convolutional neural network structure](Fig. 1)
analysis, we also repeated the whole procedure based on
the raw image data (normalized gray matter volumes)
instead of using the residuals as CNN input. Here, we set
the number of epochs to 20 due to slower convergence of
the models.

We also trained CNN models on the whole ADNI-
GO/2 sample for further evaluation. Here, we fixed
the number of epochs to 4 for the residualized data
and 8 for the raw data. These values provided the high-
est average accuracy and lowest loss in the previous
cross-validation.

**Model evaluation**

The balanced accuracy and area under the receiver oper-
ating characteristic curve (AUC) were calculated for the
independent validation samples. We report first the num-
bers for the model trained on the whole ADNI-GO/2
dataset and second the average values for the models
obtained via cross-validation.

As an internal validity benchmark, we compared
CNN model performance and group separation using
hippocampus volume, the best-established MRI
marker for Alzheimer’s disease. Automated extrac-
tion of hippocampus volume is already implemented
in commercial radiology software to aid physicians in
diagnosing dementia. We extracted total hippocam-
pus volume from the modulated and normalized MRI
scans using the Automated Anatomical Labeling (AAL)
atlas [41]. The extracted volumes were corrected for the
effects of age, sex, total intracranial volume, and mag-
netic field strength of the MRI scanner in the same way
as described above for the CNN input (see the section
“Image preparation and processing”). Here, a linear
model was estimated based on the normal controls of
the ADNI-GO/2 training sample, and then, the param-
eters were applied to the measures of all other partici-
pants and validation samples to obtain the residuals.
Subsequently, the residuals of the training sample were
entered into a receiver operating characteristic analy-
sis to obtain the AUC. The optimal threshold providing
the highest accuracy was selected based on the Youden
index. We obtained two thresholds. One for the separa-
tion of MCI and controls, which was the residual
volume of \(-0.63\) ml. That means participants with the
development of individual hippocampus volume from the
expected value (for that age, sex, total brain volume, and
magnetic field strength) below \(-0.63\) ml were classi-
cified as MCI. The other threshold for AD dementia and
controls was \(-0.95\) ml. Additionally, we repeated the
same cross-validation training/test splits as used for
CNN training to compare the variability of the derived
thresholds and performance measures.

**CNN relevance map visualization**

Relevance maps were derived from the CNN models
using the LRP algorithm [4] implemented in the Python
package iNNvestigate 1.0.9 [42]. LRP has previously been
demonstrated to yield relevance maps with high spa-
tial resolution and clinical plausibility [11, 14]. In this
approach, the final network activation scores for a given
input image are propagated back through the network
layers. LRP applies a relevance conservation principle
that means that the total amount of relevance per layer
is kept constant during the back-tracing procedure to
reduce numerical challenges that occur in other methods
[4]. Several rules exist, which apply different weighting
to positive (excitatory) and negative (inhibitory) con-
nections such that network activation for and against a
specific class can be considered differentially. Here, we
applied the so-called \(\alpha = 1, \beta = 0\) rule that only considers
positive relevance as proposed by [11, 14]. In this case,
the relevance of a network neuron \(R_j\) was calculated from
all connected neurons \(k\) in the subsequent network layer
using the formula:

\[
R_j = \sum_k \frac{a_j w_{jk}^+}{\sum_l a_l w_{lk}^+} R_k
\]  

with \(a_j\) being the activation of neuron \(j\), \(w_{jk}^+\) being the
positive weight of the connection between neurons \(j\) and \(k\),
and \(R_j\) being the relevance attributed to neuron \(k\) [5]. As
recent studies reported further improvements in LRP rele-
ance attribution [43, 44], we applied the LRP \(\alpha = 1, \beta = 0\)
composition rule that applies (3) to the convolutional
layers, and the slightly extended \(\epsilon\) rule [5] to the fully con-
nected layers. In the \(\epsilon\) rule, (3) is being extended by a small
constant term added to the denominator, i.e., \(\epsilon = 10^{-10}\)
in our case, which is expected to reduce relevance when the
activation of neuron \(k\) is weak or contradictory [5].

To facilitate model assessment and quick inspection of
relevance maps, we implemented an interactive Python
visualization application that is capable of immediate
switching between CNN models and participants. More
specifically, we used the Bokeh Visualization Library 2.2.3
(https://bokeh.org). Bokeh provides a webserver backend
and web browser frontend to directly run Python code
that dynamically generates interactive websites contain-
ning various graphical user interface components and
plots. The Bokeh web browser JavaScript libraries han-
dle the communication between the browser and server
instance and translate website user interaction into
Python function calls. In this way, we implemented vari-
ous visualization components to adjust plotting param-
eters and provide easy navigation for the 2D slice views
obtained from the 3D MRI volume.
The application is structured following a model–view–controller paradigm. An overview of implemented functions is provided in Supplementary Fig. 1. A sequence diagram illustrating function calls when selecting a new person is provided in Supplementary Fig. 2. The source code and files required to run the interactive visualization are publicly available via https://github.com/martindyrba/DeepLearningInteractiveVis.

As core functionality, we implemented the visualization in a classical 2D multi-slice window with axial, coronal, and sagittal views, cross-hair, and sliders to adjust the relevance threshold as well as minimum cluster size threshold (see Fig. 2). Here, a cluster refers to groups of adjacent voxels with high relevance above the selected relevance threshold. The cluster size is the number of voxels in this group and can be controlled in order to reduce the visual noise caused by single voxels with high relevance. Additionally, we added visual guides to improve usability, including (a) a histogram providing the distribution of cluster sizes next to the cluster size threshold slider, (b) plots visualizing the amount of positive and negative relevance per slice next to the slice selection sliders, and (c) statistical information on the currently selected cluster. Furthermore, assuming spatially normalized MRI data in MNI reference space, we added (d) atlas-based anatomical region lookup for the current cursor/cross-hair position and (e) the option to display the outline of the anatomical region to simplify visual comparison with the cluster location.

**CNN model comprehensibility and validation**

As quantitative metrics for assessing relevance map quality are still missing, we compared CNN relevance scores in the hippocampus with hippocampus volume. Here, we used the same AAL atlas hippocampus masks as for deriving hippocampus volume and applied it on the relevance maps obtained from all ADNI-GO/2 participants for each model. The sum of relevance score of each voxel inside the mask was considered as hippocampus relevance. Hippocampus relevance and volume were compared using Pearson’s correlation coefficient.

Additionally, we visually examined a large number of scans from each group to derive common relevance patterns and match them with the original MRI scans. Furthermore, we calculated mean relevance maps for each group. We also extracted the relevance for all lobes of the brain and subcortical structures to test the specificity of relevance distribution across the whole brain. These masks were defined based on the other regions included in the AAL atlas [41].
In an occlusion sensitivity analysis, we evaluated the influence of local atrophy on the prediction of the model and the derived relevance scores. Here, we slid a cube of 20 voxels = 30 mm edge size across the brain. Within the cube, we reduced the intensity of the voxel by 50%, simulating gray matter atrophy in this area. We selected a normal control participant from the DELCODE dataset without visible CNN relevance, a prediction probability for AD/MCI of 20%, and hippocampus volume residual of 0 ml, i.e., the hippocampus volume matched the reference volume expected for this person. For each position of the cube, we derived the probability of AD predicted by the model obtained from the whole ADNI-GO/2 sample. Additionally, we calculated the total amount of relevance in the scan.

Results

Group separation
The accuracy and AUC for diagnostic group separation are shown in Table 3. Additional performance measures are provided in Supplementary Table 1. The CNN reached a balanced accuracy between 75.5 and 88.3% across validation samples with an AUC between 0.756 and 0.976 for separating AD dementia and controls. For MCI vs. controls, the group separation was substantially lower with balanced accuracies between 63.1 and 75.4% and an AUC between 0.667 and 0.840. These values were only slightly better than the group separation performance of hippocampus volume (Table 3). The performance results for the raw gray matter volume data as input for the CNN are provided in Supplementary Table 2. In direct comparison to the CNN results for the residualized data, the balanced accuracies and AUC values did not show a clear difference (Table 3, Supplementary Table 2).

Model comprehensibility and relevance map visualization
The implemented web application frontend is displayed in Fig. 2. The source code is available at https://github.com/martindyrba/DeepLearningInteractiveVis and the web application can be publicly accessed at https://explanation.net/demo. In the left column, the user can select a study participant and a specific model. Below, there are controls (sliders) to adjust the thresholds for displayed relevance score, cluster size, and overlay transparency. As we used the spatially normalized MRI images as CNN input, we can directly obtain the anatomical reference location label from the automated

### Table 3: Group separation performance for hippocampus volume and the convolutional neural network models

| Sample    | Balanced accuracy (mean ± SD) | AUC              | Balanced accuracy (mean ± SD) | AUC (mean ± SD) |
|-----------|-------------------------------|------------------|-------------------------------|-----------------|
| **ADNI-GO/2** |                               |                  |                               |                 |
| MCI vs. CN | (70.0% ± 6.8%)                | (0.773 ± 0.091)  | (74.5% ± 6.2%)                | (0.785 ± 0.078) |
| AD vs. CN  | (84.4% ± 3.6%)                | (0.945 ± 0.024)  | (88.9% ± 5.3%)                | (0.949 ± 0.029) |
| MCI+ vs. CN− | (75.6% ± 7.1%)             | (0.831 ± 0.080)  | (86.7% ± 10.3%)               | (0.925 ± 0.071) |
| AD+ vs. CN− | (86.2% ± 4.2%)              | (0.954 ± 0.025)  | (94.9% ± 3.8%)                | (0.985 ± 0.017) |
| **ADNI-3** |                               |                  |                               |                 |
| MCI vs. CN | 62.8% (63.1% ± 1.4%)          | 0.683            | 63.1% (63.6% ± 1.5%)          | 0.684 (0.677 ± 0.020) |
| AD vs. CN  | 83.4% (83.4% ± 0.4%)          | 0.917            | 84.4% (81.7% ± 2.9%)          | 0.913 (0.899 ± 0.013) |
| MCI+ vs. CN− | 69.1% (69.2% ± 2.7%)       | 0.791            | 69.8% (68.3% ± 4.4%)          | 0.810 (0.742 ± 0.024) |
| AD+ vs. CN− | 83.6% (82.0% ± 1.8%)       | 0.882            | 80.2% (75.5% ± 4.2%)          | 0.830 (0.828 ± 0.028) |
| **AIBL**  |                               |                  |                               |                 |
| MCI vs. CN | 67.4% (67.6% ± 0.5%)          | 0.741            | 68.2% (67.3% ± 2.7%)          | 0.763 (0.749 ± 0.012) |
| AD vs. CN  | 84.1% (85.3% ± 1.5%)          | 0.927            | 85.0% (82.3% ± 3.0%)          | 0.950 (0.926 ± 0.007) |
| MCI+ vs. CN− | 78.5% (78.8% ± 0.9%)      | 0.874            | 75.4% (73.6% ± 3.1%)          | 0.828 (0.814 ± 0.022) |
| AD+ vs. CN− | 87.2% (89.1% ± 2.4%)        | 0.976            | 88.3% (85.3% ± 3.3%)          | 0.978 (0.958 ± 0.011) |
| **DELCODE** |                               |                  |                               |                 |
| MCI vs. CN | 69.0% (69.0% ± 9.6%)          | 0.774            | 71.0% (69.7% ± 2.6%)          | 0.775 (0.772 ± 0.017) |
| AD vs. CN  | 88.4% (86.4% ± 3.0%)          | 0.943            | 85.5% (80.5% ± 4.0%)          | 0.953 (0.938 ± 0.013) |
| MCI+ vs. CN− | 77.4% (77.8% ± 0.7%)      | 0.867            | 72.2% (74.9% ± 3.5%)          | 0.840 (0.830 ± 0.017) |
| AD+ vs. CN− | 88.2% (87.6% ± 1.8%)        | 0.954            | 83.3% (82.2% ± 4.0%)          | 0.968 (0.956 ± 0.012) |

Reported values are for the single model trained on the whole ADNI-GO/2 dataset. In parenthesis, the mean values and standard deviation for the ten models trained in the tenfold cross-validation procedure are provided to indicate the variability of the measures. Values for the ADNI-GO/2 sample (in italics) may be biased as the respective test subsamples were used to determine the optimal model during training. We still report them for better comparison of the model performance across samples.
user interactions are directly sent to the server, facilitating navigation to slices with high relevance. All cumulated relevance score per slices as visual guides to the 3D volume. We further added sliders and plots of depicted above the coronal, axial, and sagittal views of field strength) and the CNN likelihood score for AD are as covariates (age, sex, total intracranial volume, MRI cross-validation, the median correlation of total hippocampus relevance and volume showed a tendency $p < 0.001$). Across all ten models obtained using CODE validation sample are shown in Fig. 4 and those for the ADNI-GO/2 training sample in Supplementary Fig. 3. They are very similar to traditional statistical maps obtained from voxel-based morphometry, indicating the highest contribution of medial temporal brain regions, more specifically the hippocampus, amygdala, thalamus, middle temporal gyrus, and middle/posterior cingulate cortex. Also, they were highly consistent between samples (Supplementary Fig. 3).

The occlusion sensitivity analysis also showed identical brain regions’ atrophy to contribute to the model’s decision (Fig. 5). Interestingly, the occlusion relevance maps showed a ring structure around the most contributing brain areas, indicating that relevance was highest when the occluded area just touched the salient regions, leading to a thinning-like shape of the gray matter.

The correlation of individual DELCODE participants’ hippocampus relevance score and hippocampus volume for the model trained on the whole ADNI-GO/2 dataset is displayed in Fig. 6. For this model, the correlation was $r = −0.87$ for bilateral hippocampus volume ($p < 0.001$). Across all ten models obtained using cross-validation, the median correlation of total hippocampus relevance and volume was $r = −0.84$ with a range of $−0.88$ and $−0.44$ (all with $p < 0.001$). Cross-validation models with higher correlation between hippocampus relevance and volume showed a tendency for better AUC values for MCI vs. controls ($r = 0.61$, $p = 0.059$). To test whether hippocampus volume and relevance measures were specific to the hippocampus, we also compared the correlation between hippocampus volume and other regions’ and whole-brain relevance. Here, the correlations were lower, with $r = −0.62$ ($p < 0.001$) between hippocampus volume and whole-brain relevance. More detailed results are provided as a correlation matrix in Supplementary Fig. 4.

Discussion

Neural network comprehensibility

We have presented a CNN framework and interactive visualization application for obtaining class-specific relevance maps for disease detection in MRI scans, yielding human-interpretable and clinically plausible visualizations of key features for image discrimination. To date, most CNN studies focus on model development and optimization, which are undoubtedly important tasks and there are still several challenges to tackle. However, as black-box models, it is typically not feasible to judge, why a CNN fails or which image features drive a particular decision of the network. This gap might be closed with the use of novel visualization algorithms such as LRP [4] and deep Taylor decomposition [5]. In our application, LRP relevance maps provided a useful tool for model inspection to reveal the brain regions which contributed most to the decision process encoded by the neural network models.

Currently, there is no ground truth information for relevance maps, and there are no appropriate methods available to quantify relevance map quality. Samek and colleagues [45] proposed the information-theoretic measures relevance map entropy and complexity, which mainly characterize the scatter or smoothness of images. Furthermore, adapted from classical neural network sensitivity analysis, they assessed the robustness of relevance maps using perturbation testing where small image patches were replaced by random noise, which was also applied in [46]. Already for 2D data, this method is computationally very expensive and only practical for a limited number of input images. Instead of adding random noise, we simulated gray matter atrophy by lowering the image intensities by 50% in a cube-shaped area. As visible from Fig. 5, the brain areas contributing to the model’s AD probability nicely matched the areas shown in the mean relevance maps (Fig. 4). Notably, the ring-shaped increase in relevance around the salient regions (Fig. 5, bottom) indicates that the model is sensitive to intensity jumps occurring when the occlusion cube touches the borderline of those regions. Most probably, this means that the model was more sensitive to thinning patterns of gray matter than to equally distributed volume reduction. However, our findings have to be seen as preliminary, as we only assessed this analysis in one normal control...
Fig. 3 Example relevance maps obtained for different people. Top row: Alzheimer’s dementia patients, middle row: patients with mild cognitive impairment, bottom row: cognitively normal controls.
participant due to the computational effort, and therefore, it requires more extensive research in future studies.

Based on the extensive knowledge about the effect of Alzheimer’s disease on brain volume as presented in T1-weighted MRI scans [15, 16], we selected a direct quantitative comparison of relevance maps with hippocampus volume as a validation method. Here, we obtained very high correlations between hippocampus relevance scores and volume (median correlation $r = -0.84$), underlining the clinical plausibility of learnt patterns to differentiate AD and MCI patients from controls. In addition, visual inspection of relevance maps also revealed several other clusters with gray matter atrophy in the individual participants’ images that contributed to the decision of the CNN (Figs. 2 and 3). Böhle and colleagues [14] proposed an atlas-based aggregation of CNN relevance maps to be used as “disease fingerprints” and to enable a quick comparison between patients and controls, a concept that has also been proposed previously for differential diagnosis of dementia based on heterogeneous clinical data and other machine learning models [47, 48].

Notably, the CNN models presented here were solely based on the combinations of input images with their corresponding diagnostic labels to determine which brain features were diagnostically relevant. Traditionally, extensive clinical experience is required to define relevant features (e.g., hippocampus volume) that discriminate between a clinical population (here: AD, MCI) and a healthy control group. Also, typically, only few predetermined parameters are used (e.g., hippocampus volume or medial temporal lobe atrophy score [15, 16]). Our results demonstrate that the combination of CNN and relevance map approaches constitutes a promising tool for improving the utility of CNN in the classification of MRIs of patients with suspected AD in a clinical context. By referring back to the relevance maps, trained clinicians will be enabled to compare classification results to comprehensible features visible in the relevance images and thereby more readily interpret the classification results in clinically ambiguous situations. Perspectively, the relevance map approach might also provide a helpful tool to reveal features for more complex diagnostic challenges such as differential diagnosis between various types of dementia.
for instance the differentiation between AD, frontotemporal dementia, and dementia with Lewy bodies.

**CNN performance**

As expected, CNN-based classification reached an excellent $AUC \geq 0.91$ for the group separation of AD compared to controls but a substantially lower accuracy for group separation between MCI and controls ($AUC \approx 0.74$, Table 3). When restricting the classification to amyloid-positive MCI versus amyloid-negative controls, group separation improved to $AUC = 0.84$ in DELCODE, highlighting the heterogeneity of MCI as a diagnostic entity and the importance of biomarker stratification [1, 2]. In summary, these numbers are also reflected by the recent CNN literature as shown in Table 1. Notably, [27] reported several limitations and issues in the performance evaluation of some other CNN papers, such that it is not easy to finally conclude on the group separation capabilities of the CNN models in realistic settings. To overcome such challenges, we validated the models on three large independent cohorts (Table 3), providing strong evidence for their generalizability and for the robustness of our CNN approach.

To put the CNN model performance into perspective, we compared the accuracy of the CNN models with the accuracy achieved by assessing hippocampus volume, the key clinical MRI marker for neurodegeneration in Alzheimer’s disease [1, 2]. Interestingly, there were only minor differences in the achieved $AUC$ values across all samples (Table 3). The MCI group of the ADNI-3 sample, which yielded the worst group separation of all samples ($AUC = 0.68$), was actually the group with the largest average hippocampus volumes and, therefore, the lowest group difference compared to the controls (Table 2). Obviously, our results here indicate a limited value of using CNN models instead of traditional volumetric markers for the detection of Alzheimer’s dementia and mild cognitive impairment. Previous MRI CNN papers have not reported the baseline accuracy reached by hippocampus volume for comparison. However, as noted above, CNNs might provide a useful tool to automatically derive discriminative features for complex diagnostic tasks where

![Fig. 5 Results from the occlusion sensitivity analysis. A gray matter volume loss of 50% was simulated in a cube of 30-mm edge length. Each voxel encodes the derived values when centering the cube at that position. Top: probability of AD for the areas with simulated atrophy. Bottom: total sum of image relevance depending on simulated atrophy. Numbers indicate the y-axis slice coordinates in MNI reference space.](image-url)
Clear clinical criteria are still missing, for instance for the differential diagnosis between various types of dementia.

Limitations
As already mentioned above, visual inspection of relevance maps also revealed several other regions with gray matter atrophy in the individual participants’ images that contributed to the decision of the CNN. These additional regions were not further assessed, as a priori knowledge regarding their diagnostic value is still under debate in the scientific community [1, 2]. Also, we did not perform a three-way classification between AD dementia, MCI, and CN due to the limited availability of cases for training. Additionally, MCI itself is a heterogeneous diagnostic entity [1, 2]. Here, all the studies involved in our analysis tried to increase the likelihood of underlying Alzheimer’s pathology by focusing on MCI patients with memory impairment. But markers of amyloid-beta pathology were only available for a subset of participants such that we could not stratify by amyloid status for the training of the CNN models. However, we optionally applied this stratification for the validation of the CNN performances to improve the diagnostic confidence.

Future prospects
Several studies focused on CNN models for the integration of multimodal imaging data, e.g., MRI and fluorodeoxyglucose (FDG)-PET [17–19], or heterogeneous clinical data [49]. Here, it will be beneficial, to directly include the variables we used as covariates (such as age and sex) as input to the CNN model rather than performing the variance reduction directly on the input data before applying the model. In this context, relevance mapping visualization approaches need to be developed that allow for a direct comparison of the relevance magnitude for images and clinical variables simultaneously. Another aspect is the automated generation of textual descriptions and diagnostic explanations from images [50–52]. Given the recent technical progress, we suggest that the approach is now ready for interdisciplinary exchange to assess how clinicians can benefit from CNN assistance in their diagnostic workup, and which requirements must be met to increase clinical utility. Beyond the technical challenges, regulatory and ethical aspects and caveats must be carefully considered when introducing CNN as part of clinical decision support systems and medical software—and the discussion of these issues has just recently begun [53, 54].
Conclusion
We presented a framework for obtaining diagnostic relevance maps from CNN models to improve model comprehensibility. These relevance maps have revealed reproducible and clinically plausible atrophy patterns in AD and MCI patients, with a high correlation with the well-established MRI marker of hippocampus volume. The implemented web application allows a quick and versatile inspection of brain regions with a high relevance score in individuals. With the increased comprehensibility of CNNs provided by the relevance maps, the data-driven and hypothesis-free CNN modeling approach might provide a useful tool to aid differential diagnosis of dementia and other neurodegenerative diseases, where fine-grained knowledge on discriminating brain alterations is still missing.

Supplementary Information
The online version contains supplementary material available at https://doi.org/10.1186/s13195-021-00924-2.

Additional file 1: Supplementary Table 1. Group separation performance for hippocampus volume and the convolutional neural network models for residualized data (extended).

Additional file 2: Supplementary Table 2. Group separation performance for hippocampus volume and the convolutional neural network models for raw input data.

Additional file 3: Supplementary Figure 1. UML diagram of the interactive visualization application.

Additional file 4: Supplementary Figure 2. Sequence diagram of function calls when selecting a new person.

Additional file 5: Supplementary Figure 3. Comparison of mean relevance maps between samples.

Additional file 6: Supplementary Figure 4. Correlation matrix of hippocampus volume (residualized) and several brain regions’ relevance scores for DELCODE participants and the model trained on the whole ADNI-GO/2 dataset.

Acknowledgements
The data samples were provided by the DELCODE study group of the Clinical Research Unit of the German Center for Neurodegenerative Diseases (DZNE). Details and participating sites can be found at www.dzne.de/en/research/studies/clinical-studies/delcode. The DELCODE study was supported by Max Delbrück Center for Molecular Medicine in the Helmholtz Association (MDC), Berlin, Center for Cognitive Neuroscience Berlin (CCNB) at Freie Universität Berlin, Bernstein Center for Computational Neuroscience (BCCN), Berlin, Core Facility MRI-Research in Neurosciences, University Medical Center Goettingen, Institute for Clinical Radiology, Ludwig Maximilian University, Munich; Institute of Diagnostic and Interventional Radiology, Pediatric Radiology and Neuroradiology, Rostock University Medical Center, and Magnetic Resonance research center, University HospitalTuebingen.

Data collection and sharing for this project was funded by the Alzheimer’s Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904). 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; Cerecor 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.; Meiro 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. A complete listing of ADNI investigators can be found at http://adni.loni.ucsd.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf. ADNI researchers are listed at aibl.csiro.au.

Authors’ contributions
MD: conceptualization, methodology, data curation and processing, coding and software development, visualization, and writing of the original draft.
MH: coding and software development and visualization. ST: conceptualization, methodology, data collection and curation, writing, review and editing, supervision, and clinical validation. All others: data acquisition, collection and curation, substantial intellectual contribution on study design and methodology, review and editing, and clinical validation. All authors read and approved the final manuscript.

Funding
This study was supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation), project ID 454834942, funding code DY151/2-1. Open Access funding enabled and organized by Projekt DEAL.

Availability of data and materials
Data used for training/evaluation of the models is available from the respective initiatives (ADNI: http://adni.loni.usc.edu/data-samples/access-data, AIBL: https://aibl.csiro.au, DELCODE: https://www.dzne.de/en/research/studies/clinical-studies/delcode). The source code, a demo dataset, the trained CNN models, and all additional files required to run the interactive visualization are publicly available at GitHub: https://github.com/martindyrba/DeepLearningInteractiveVIs.

Declarations
Ethics approval and consent to participate
Data collecting within ADNI and AIBL was approved by participating institutions. See https://adni.loni.usc.edu and https://aibl.csiro.au for details. The DELCODE study was approved by participating institutions. See [35] for details. All study participants or their representatives provided written informed consent to participate in the respective studies and also agreed to sharing of their data. The retrospective analysis, study design, and interactive visualization of relevance maps were approved by the internal review board of the Rostock University Medical Center, reference number A 2020-0182.

Consent for publication
Not applicable.
Competing interests
The authors declare that they have no competing interests.

Author details
1German Center for Neurodegenerative Diseases (DZNE), Rostock, Germany. 2Institute of Visual and Analytic Computing, University of Rostock, Rostock, Germany. 3German Center for Neurodegenerative Diseases (DZNE), Berlin, Germany. 4Department of Psychiatry and Psychotherapy, Charité – Universitätsmedizin Berlin, Campus Charité Mitte, Berlin, Germany. 5German Center for Neurodegenerative Diseases (DZNE), Bonn, Germany. 6Department for Neurodegenerative Diseases and Geriatric Psychiatry, University Hospital Bonn, Bonn, Germany. 7German Center for Neurodegenerative Diseases (DZNE), Munich, Germany. 8Institute for Stroke and Dementia Research (ISD), University Hospital, Ludwig Maximilian University, Munich, Germany. 9Institute of Diagnostic and Interventional Radiology, Pediatric Radiology and Neuroradiology, Rostock University Medical Center, Rostock, Germany. 10MRI-Research in Neurosciences, Department of Cognitive Neurology, Georg-August-University, Goettingen, Germany. 11Department for Neurodegenerative Diseases (DZNE), Magdeburg, Germany. 12Institute of Cognitive Neurology and Dementia Research (IKND), Otto-von-Guericke University, Magdeburg, Germany. 13Bernstein Center for Computational Neuroscience, Berlin, Germany. 14Department of Psychiatry and Psychotherapy, Charité – Universitätsmedizin Berlin, Campus Benjamin Franklin, Berlin, Germany. 15Department of Psychosomatic Medicine, Heinrich-Heine-University Medical Center, Düsseldorf, Germany. 16German Center for Neurodegenerative Diseases (DZNE), Magdeburg, Germany. 17Section for Dementia Research, Hertie Institute for Clinical Brain Research, Tübingen, Germany. 18Department of Psychiatry and Psychotherapy, University of Tübingen, Tübingen, Germany. 19Department of Psychiatry, Medical Faculty, University of Cologne, Cologne, Germany. 20Department of Psychiatry and Psychotherapy, Otto-von-Guericke University, Magdeburg, Germany. 21Systems Neurophysiology, Department of Biology, Darmstadt University of Technology, Darmstadt, Germany. 22Department of Psychiatry and Psychotherapy, University Hospital, Ludwig Maximilian University, Munich, Germany. 23Munich Cluster for Systems Neurology (SyNergy), Ludwig Maximilian University, Munich, Germany. 24Ageing Epidemiology Research Unit (AGE), School of Public Health, Imperial College London, London, UK. 25Department of Psychiatry and Psychotherapy, Klinikum rechts der Isar, Technical University Munich, Munich, Germany. 26Department of Biomedical Magnetic Resonance, University of Tübingen, Tübingen, Germany. 27German Center for Neurodegenerative Diseases (DZNE), Goettingen, Germany. 28Department of Psychiatry and Psychotherapy, University Medical Center Goettingen, Goettingen, Germany. 29Leibniz Institute for Neurobiology, Magdeburg, Germany. 30Institute of Clinical Radiology, Ludwig Maximilian University, Munich, Germany. 31Institute for Molecular Genetics, Otto-von-Guericke University, Magdeburg, Germany. 32Department of Neurology, University Hospital Bonn, Bonn, Germany. 33Neurosciences and Signaling Group, Institute of Biomedicine (iBIMED), Department of Medical Sciences, University of Aveiro, Aveiro, Portugal. 34Excellence Cluster on Cellular Stress Responses in Aging-Associated Diseases (CECAD), University of Cologne, Cologne, Germany.

Received: 2 March 2021 Accepted: 25 October 2021

Page 17 of 18

References
1. Jack CR, Albert MS, Knopman DS, McKhann GM, Frisoni GB, Jagust EJ, et al. Introduction to the recommendations from the National Institute on Aging-Alzheimer’s Association workgroups on diagnostic guidelines for Alzheimer’s disease. Alzheimers Dement. 2011;7(3):257–62.
2. Dubois B, Feldman HH, Jacova C, Hampel H, Molinuevo JL, Blennow K, et al. Advancing research diagnostic criteria for Alzheimer’s disease: the IWG-2 criteria. Lancet Neurol. 2014;13(6):614–29.
3. Vemuri P, Fields J, Peter J, Kloppel S. Cognitive interventions in Alzheimer’s and Parkinson’s disease. Curr Opin Neurol. 2016;29(4):405–11.
4. Bach S, Binder A, Montavon G, Klauschen F, Müller K-R, Samek W. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PLoS One. 2015;10(7).
5. Montavon G, Samek W, Müller K-R. Methods for interpreting and understanding deep neural networks. Digital Signal Process. 2018;73:1–15.
6. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: visual explanations from deep networks via gradient-based localization. In: 2017 IEEE International Conference on Computer Vision (ICCV). 2017:618–626.
7. Zeiler MD, Fergus R. Visualizing and understanding convolutional networks. In: Computer Vision – ECCV 2014. 2014:818–833.
8. Thibade-Sutter E, Collot O, Dormont D, Burges N, Landman BA, Glimm J. Visualisation approach to assess the robustness of neural networks for medical image classification. In: Medical Imaging 2020: Image Processing. 2020.
9. Ribeiro MT, Singh S, Guestrin C. “Why should I trust you?” In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016:1135–1144.
10. Alber M. Software and application patterns for explanation methods. In: Explainable AI: interpreting, explaining and visualizing deep learning. 2019:399–433.
11. Dyrba M, Pallath AH, Marzbani EN. Comparison of CNN visualization methods to aid model interpretability for detecting Alzheimer’s disease. In: Bildverarbeitung für die Medizin. 2020:307–312.
12. Etel F, Ritter K. Testing the robustness of attribution methods for convolutional neural networks in MRI-based Alzheimer’s disease classification. In: Interpretability of machine intelligence in medical image computing and multimodal learning for clinical decision support. 2019:3–11.
13. Springenberg JT, Dosovitskiy A, Brox T, Riedmiller M. Strengthening simplicity: the all convolutional net. In: 3rd International Conference on Learning Representations, ICLR 2015, Workshop Track Proceedings. Edited by Bengio Y, LeCun Y.
14. Böhle M, Etel F, Weygandt M, Ritter K. Layer-wise relevance propagation for explaining deep neural network decisions in MRI-based Alzheimer’s disease classification. Front Aging Neurosci. 2019;11:194.
15. Scheltens P, Leys D, Barkhof F, Huglo D, Weinstein HC, Vermeersch P, et al. Atrophy of medial temporal lobes on MRI in “probable” Alzheimer’s disease and normal ageing: diagnostic value and neuropsychological correlates. J Neurol Neurosurg Psychiatry. 1992;55(10):967–72.
16. Teipel S, Drzezga A, Grothe MJ, Barthel H, Chételat G, Schiff N, et al. Multimodal imaging in Alzheimer’s disease: validity and usefulness for early detection. Lancet Neurol. 2015;14(10):1037–53.
17. Suk H-S, Lee S-W, Shen D. Hierarchical feature representation and multimodal fusion with deep learning for AD/MCI diagnosis. NeuroImage. 2014;101:569–82.
18. Li F, Tran L, Thung K-H, Ji S, Shen D, Li J. A robust deep model for improved classification of AD/MCI patients. IEEE J Biomed Health Inform. 2015;19(5):1610–6.
19. Ortz A, Munilla J, Goritz JM, Ramírez J. Ensembles of deep learning architectures for the early diagnosis of the Alzheimer’s disease. Int J Neural Syst. 2016;26(7):1650025.
20. Aderghal K, Khvostikov A, Krylov A, Benaïs-Pineau J, Afdel K, Catheline G. Classification of Alzheimer disease on imaging modalities with deep CNNs using cross-modal transfer learning. In: 2018 IEEE 31st International Symposium on Computer-Based Medical Systems (CBMS). 2018:345–350.
21. Liu M, Cheng D, Yan W. Classification of Alzheimer’s disease by combination of convolutional and recurrent neural networks using FDG-PET images. Front Neuroinfomat. 2018;12.
22. Liu M, Zhang J, Nie D, Yap P-T, Shen D. Anatomical landmark based deep feature representation for MR images in brain disease diagnosis. IEEE J Biomed Health Inform. 2018;22(5):1676–86.
23. Lin W, Tong T, Gao Q, Guo D, Du X, Yang Y, et al. Convolutional neural networks-based MRI image analysis for the Alzheimer’s disease prediction from mild cognitive impairment. Front Neurosci. 2018;12.
24. Li H, Habes M, Wolk DA, Fan Y. A deep learning model for early prediction of Alzheimer’s disease dementia based on hippocampal magnetic resonance imaging data. Alzheimers Dement. 2019;15(8):1059–70.
25. Lin C, Liu M, Zhang J, Shen D. Hierarchical fully convolutional network for joint atrophy localization and Alzheimer’s disease diagnosis using structural MRI. IEEE Trans Pattern Anal Mach Intell. 2020;42(4):880–93.
26. Qiu S, Josi PS, Miller MI, Xue C, Zhou X, Karjadi C, et al. Development and validation of an interpretable deep learning framework for Alzheimer’s disease classification. Brain. 2020;143(6):1920–33.
27. Wen J, Thibade-Sutre E, Diaz-Melo M, Samper-González J, Routier A, Bottani S, et al. Convolutional neural networks for classification of Alzheimer’s disease: overview and reproducible evaluation. Med Image Anal. 2020;63.
28. Jo T, Nho K, Risacher SL, Saykin AJ. Deep learning detection of informative features in tau PET for Alzheimer's disease classification. BMC Bioinformatics. 2020;21(21).

29. Krizhevsky A, Sutskever I, Hinton GE. ImageNet Classification with deep convolutional neural networks. In: Advances in neural information processing systems 25. Edited by Pereira F, Burges CJC, Bottou L, Weinberger KQ. Curran Associates, Inc; 2012:1097–1105.

30. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. In: 3rd International Conference on Learning Representations (ICLR 2015). Edited by Bengio Y; LeCun Y; 2015.

31. Landau SM, Mintun MA, Joshi AD, Koepp RA, Petersen RC, Aisen PS, et al. Amyloid deposition, hypometabolism, and longitudinal cognitive decline. Ann Neurol. 2012;72(4):578–86.

32. Klunk WE, Koepp RA, Price JC, Benzinger TL, Devous MD, Jagust WJ, et al. The Centiloid Project: standardizing quantitative amyloid plaque estimation by PET. Alzheimers Dement. 2015;11(1):1–15.e14.

33. Navitsky M, Joshi AD, Kennedy I, Klunk WE, Rowe CC, Wong DF, et al. Standardization of amyloid quantitation with florbetapir standardized uptake value ratios to the Centiloid scale. Alzheimers Dement. 2018;14(12):1565–71.

34. Battle MR, Pillay LC, Rowe CC, Wong DF, Eustace KE, Catak C, et al. Amyloid deposition in Alzheimer’s disease correspond to dissociable age, sex, andAPOEε4 effects on memory, brain structure, and β‑amyloid across the adult life span. JAMA Neurol. 2015;72(5).

35. Dima D, Modabbernia A, Papachristou E, Doucet GE, Agartz I, Aghajani M, Kurth F, Gaser C, Luders E. A 12‑step user guide for analyzing voxel‑wise image recognition. In: 3rd International Conference on Learning Representations (ICLR 2015). Edited by Bengio Y, LeCun Y; 2015.

36. Jessen F, Spottke A, Boecker H, Brosseron F, Buerger K, Catak C, et al. The Centiloid Project: standardizing quantitative amyloid plaque estimation by PET. Alzheimers Dement. 2015;11(1):1–15.e14.

37. Alber M, Lapuschkin S, Seegerer P, Hägele M, Schütt KT, Montavon G, et al. Sanity checks for saliency maps. In: Proceedings of the 32nd International Conference on Neural Information Processing Systems (NIPS’18). Red Hook: Curran Associates Inc; 2018. p. 9525–36.

38. Kohlbrenner M, Bauer A, Nakajima S, Binder A, Samek W, Lapuschkin S. Towards best practice in explaining neural network decisions with LRP. In: 2020 International Joint Conference on Neural Networks (IJCNN). 2020: 1–7.

39. Tzourio‑Mazoyer N, Landeau B, Papathanassiou D, Crivello F, Etard O, Delcroix N, et al. Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single‑subject brain. Neuroimage. 2002;15(1):273–89.

40. Alber M, Lapuschkin S, Seegerer P, Hägele M, Schütt KT, Montavon G, et al. iNNvestigate neural networks! J Mach Learn Res. 2019;20:1–8.

41. Krieger A, Bajwa MN, Braun SA, Malik M, Dengel A, Ahmed S. On interpretability of deep learning based skin lesion classifiers using concept activation vectors. In: International Joint Conference on Neural Networks (IJCNN–2020), July 19–24, Glasgow, United Kingdom. IEEE; 2020.

42. Adebayo J, Gilmer J, Muelly M, Goodfellow I, Hardt M, Kim B. Sanity checks for saliency maps. In: Proceedings of the 32nd International Conference on Neural Information Processing Systems (NIPS’18). Red Hook: Curran Associates Inc; 2018. p. 9525–36.

43. Kohlbrenner M, Bauer A, Nakajima S, Binder A, Samek W, Lapuschkin S. Towards best practice in explaining neural network decisions with LRP. In: 2020 International Joint Conference on Neural Networks (IJCNN). 2020: 1–7.

44. Sixt L, Granz M, Landgraf T. When explanations lie: why many modified BP attributions fail. In: Proceedings of the 37th International Conference on Machine Learning: Proceedings of Machine Learning Research: Edited by Hal D, Ill, Aarti S. PMLR 2020: 9046‑‑9057.

45. Zhang Z, Xie Y, Xing F, McGough M, Yang L. MDNet: a semantically and visually interpretable medical image diagnosis network. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2017: 6428–6436.

46. Lucieri A, Bajwa MN, Braun SA, Malik M, Dengel A, Ahmed S. On interpretability of deep learning based skin lesion classifiers using concept activation vectors. In: International Joint Conference on Neural Networks (IJCNN–2020), July 19–24, Glasgow, United Kingdom. IEEE; 2020.

47. Chronic and acute effects of tococa on cognitive function in healthy older adults: a randomized controlled trial. JAMA Neurol. 2015;72(5).

48. Alber M, Lapuschkin S, Seegerer P, Hägele M, Schütt KT, Montavon G, et al. Sanity checks for saliency maps. In: Proceedings of the 32nd International Conference on Neural Information Processing Systems (NIPS’18). Red Hook: Curran Associates Inc; 2018. p. 9525–36.

49. Candemir S, Nguyen XV, Prevedello LM, Bigelow MT, White RD, Erdal BS. Neuroimaging Initiative AsD: predicting rate of cognitive decline at baseline using a deep neural network with multilayer analysis. J Med Imaging. 2020;7(04).

50. Jing B, Xie X, Xing E. On the automatic generation of medical imaging reports. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2018: 2577-2586.

Publisher’s Note
Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.