scientific reports

Check for updates

OPEN

Interpretable brain disease classification and relevance-guided deep learning

Christian Tinauer[1], Stefan Heber[1], Lukas Pirpamer[1,2], Anna Damulina[1], Reinhold Schmidt[1], Rudolf Stollberger[3,4], Stefan Ropele[1,4] & Christian Langkammer[1,4]✉

Deep neural networks are increasingly used for neurological disease classification by MRI, but the networks' decisions are not easily interpretable by humans. Heat mapping by deep Taylor decomposition revealed that (potentially misleading) image features even outside of the brain tissue are crucial for the classifier's decision. We propose a regularization technique to train convolutional neural network (CNN) classifiers utilizing relevance-guided heat maps calculated online during training. The method was applied using T1-weighted MR images from 128 subjects with Alzheimer's disease (mean age = 71.9 ± 8.5 years) and 290 control subjects (mean age = 71.3 ± 6.4 years). The developed relevance-guided framework achieves higher classification accuracies than conventional CNNs but more importantly, it relies on less but more relevant and physiological plausible voxels within brain tissue. Additionally, preprocessing effects from skull stripping and registration are mitigated. With the interpretability of the decision mechanisms underlying CNNs, these results challenge the notion that unprocessed T1-weighted brain MR images in standard CNNs yield higher classification accuracy in Alzheimer's disease than solely atrophy.

Alzheimer's disease (AD) is the most common form of dementia with about 50 million patients and a substantial burden for our healthcare systems, caregivers and next of kin[1]. While postmortem diagnosis can be obtained from the histological examination of tissue samples from affected anatomical regions[2,3], in vivo diagnosis is hampered by clinical symptom similarities and its accuracy is rather low (71–87% sensitivity and 44–71% specificity)[4].

In addition to clinical and neuropsychological tests, medical imaging is increasingly used to strengthen diagnosis by PET imaging ligands to amyloid-β and tau proteins combined with MRI. Recently revised diagnosis criteria for AD are clinical-biological and require both clinical phenotype and biomarker evidence (Aβ or tau) of AD[5]. Although the presence of extracellular neuritic Aβ plaques is part of several diagnosis criteria their clinical value is discussed controversially, whereas selective tau ligands do reflect clinical severity and memory impairment and also serve for in-vivo Braak-staging[6]. Based on imaging tau pathology, recent fascinating data-driven work found that tau-PET can be used to identify four spatiotemporal phenotypes which exhibit different clinical profiles and longitudinal outcomes and thus opens an avenue for personalized treatment[7]. However, AD has a long prodromal and asymptomatic inflammatory phase where radioactive PET tracers cannot be used as a means for its prognosis in a healthy population. Because pathological changes are occurring decades before initial clinical manifestations, early biomarkers in a broad population might be obtained best by MRI, where volumetry and especially hippocampal atrophy are presently used as imaging markers[8–10].

Deep learning is omnipresent in medical imaging, including image reconstruction[11], segmentation[12], and classification[13,14]. Convolutional neural networks (CNNs) are utilized for neurological disease classification[15–17] and regression[18] in prevalent neurological disorders such as Alzheimer's disease[14,19–21], Parkinson's disease[22] and multiple sclerosis[23].

Despite their improved performance, those models are generally not easily interpretable by humans and deep neural networks (DNNs) are mostly seen as black boxes where data in combination with extensive learning efforts yields decisions[24]. One striking example of misguided feature extraction of DNNs is described in[25], where secondary photo watermarks identified horses better than the actual animal print. In the context of brain MRI it has been shown that learned features for age estimation are influenced by the applied registration type (linear vs. nonlinear)[18]. However, no systematic investigation of the preprocessing of brain MR images for disease

[1]Department of Neurology, Medical University of Graz, Graz, Austria. [2]Medical Image Analysis Center (MIAC) and Department of Biomedical Engineering, University of Basel, Basel, Switzerland. [3]Institute of Biomedical Imaging, Graz University of Technology, Graz, Austria. [4]BioTechMed-Graz, Graz, Austria. ✉email: christian.langkammer@medunigraz.at

classification with CNNs has been conducted, but the studies[19,20,23] aimed at explaining their applied classifier. Preprocessing is a crucial step, with skull stripping (brain extraction) creating artificial edges and interpolation and regridding necessary for registration. CNNs can incorporate these newly introduced features during training and base their classification results thereon.

Medical imaging has high legal requirements as e.g. the EU's General Data Protection Regulation (GDPR) explicitly requires the right to explanation for users subjected to decisions of an automated processing system[26] and the US are endorsing the OECD AI Principles of transparency and explainability[27]. Consequently, medical decision-supporting algorithms require verifying that this is not the result of exploiting data artifacts and that the high accuracy of classification decisions are explainable to avoid biased results[25,28]. In the present work we used heat (or saliency) mapping, which is enabling perceptive interpretability to explain a classification result in terms of maps overlaid on the input[29]. Regions in the input image contributing most to the classification result are highlighted in the heat map. From several methods currently available generating heat maps[30–34], we based our proposed method on the deep Taylor decomposition (DTD) method[35] which is a special case of layer-wise relevance propagation (LRP)[36]. LRP, has a solid theoretical framework, has been extensively validated[37,38] and can be efficiently implemented, enabling online heat map generation during training.

Besides indications from aforementioned studies, our experiments on Alzheimer's disease classification showed that CNNs might learn from (misleading) features outside the parenchyma or features introduced by the skull stripping algorithm. Thus, besides investigating how preprocessing steps including registration and skull stripping identify relevant features, we additionally present a novel relevance guided algorithm, mitigating the necessity and impact of skull stripping for classification of brain diseases. Based on its implementation this is referred to as Graz$^+$ technique (guided relevance by adaptive z^+-rule). In summary, the specific contributions of this work are:

- CNN-based disease classification in a cohort of 128 patients with AD and 290 age-matched normal controls.
- Using subject-level 3D T1-based MR image data, differently preprocessed regarding registration and skull stripping.
- Graz$^+$ technique: A relevance-guided regularization technique for CNN classifiers to mitigate the impact of MRI preprocessing.
- Making the framework's source code freely available for reproducibility of the presented results at www.neuroimaging.at/explainable-ai.

Methods

Subjects.

Inclusion criteria for all participants was a diagnosis of probable or possible AD according to the NINCDS-ADRDA criteria[39] and a complete MRI and study protocol as described in detail in[40]. The healthy control (HC) group was selected from participants of a study in community-dwelling individuals. These volunteers were randomly selected from the community register, had a normal neurological status, and were without cerebrovascular attacks and dementia as previously described[41]. This study was approved by the ethics committee of the Medical University of Graz (IRB00002556) and signed written informed consent was obtained from all study participants or their caregivers. The trial protocol for this prospective study was registered at the National Library of Medicine (trial identification number: NCT02752750). All methods were performed in accordance with the relevant guidelines and regulations.

MR imaging.

Patients and controls were scanned using a consistent MRI protocol at 3 Tesla (Magnetom TimTrio; Syngo MR B17; Siemens Healthineers, Erlangen, Germany) using a 12-channel phased-array head coil. Structural imaging included a T1-weighted 3D MPRAGE sequence with 1 mm isotropic resolution (TR/TE/TI/FA = 1900 ms/2.19 ms/900 ms/9°, matrix = 176 × 224 × 256).

Data selection.

Totally 132 patients with probable AD with 295 scans[40] and 381 controls with 514 scans from an ongoing community dwelling study[41] were included in this retrospective study. From patients we excluded 12 MRIs because T1-weighted images were not available, and 14 scans because the image matrix was differently sized. Similarly, from controls we excluded 13 MRIs because of missing T1-weighted images, and 17 MRIs because of different image matrix sizes. Age-matching was achieved by excluding 5 scans of patients and 106 scans of controls, yielding 264 T1-weighted images from 128 patients with probable AD (mean age = 71.9 ± 8.5 years) and 378 MRIs from 290 healthy controls (mean age = 71.3 ± 6.4 years) for the subsequent deep learning analysis.

Preprocessing.

Brain masks from T1-weighted MRIs were obtained using BET from FSL 6.0.3 with bias field/neck cleanup enabled and a fractional intensity threshold of 0.35[42]. T1-weighted images were registered to the MNI152 T1 template (A) affinely, using FSL flirt with 6 degrees of freedom and a correlation ratio based cost function, and (B) nonlinearly, using FSL fnirt with the *T1_2_MNI152_2mm* configuration. Bias field correction was not applied to the T1-weighted input images, but image intensities were divided by 450 to obtain image intensity ranges between 0 and 1 to speed up the network trainings.

Attention mask.

Our relevance-guided method is preconditioned by binary attention masks. We used entire brain masks obtained by FSL-BET to focus the classifiers to the intracranial volume.

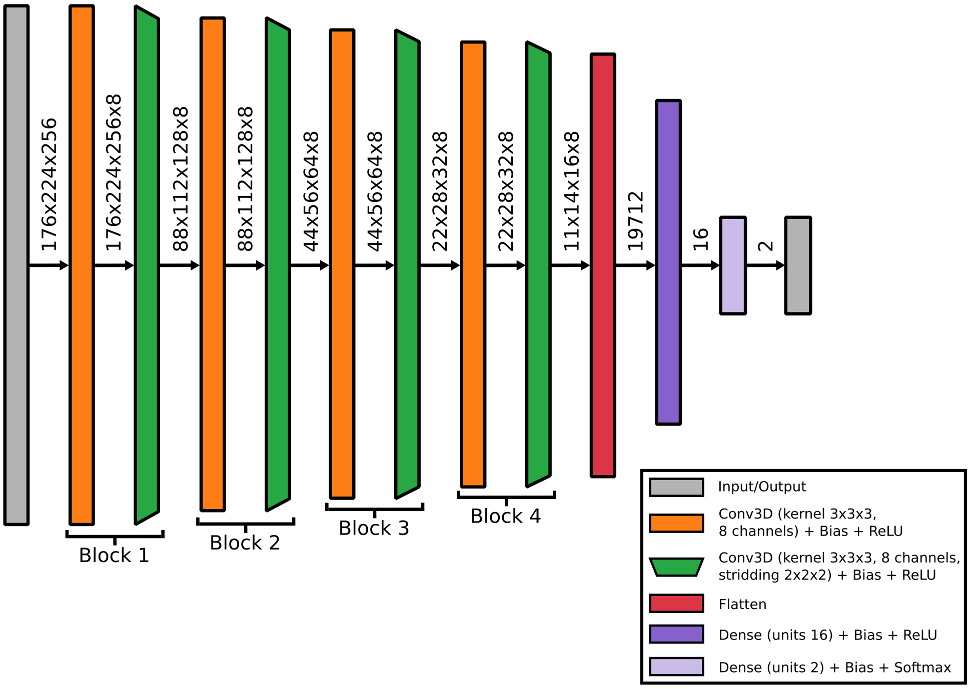
176x224x256
176x224x256x8
88x112x128x8
88x112x128x8
44x56x64x8
44x56x64x8
22x28x32x8
22x28x32x8
11x14x16x8
19712
16
2
Block 1
Block 2
Block 3
Block 4
Input/Output
Conv3D (kernel 3x3x3, 8 channels) + Bias + ReLU
Conv3D (kernel 3x3x3, 8 channels, stridding 2x2x2) + Bias + ReLU
Flatten
Dense (units 16) + Bias + ReLU
Dense (units 2) + Bias + Softmax

Figure 1. Structure of the 3D classifier network combining a single convolutional layer (kernel 3 × 3 × 3, 8 channels) with a down-convolutional layer (kernel 3 × 3 × 3, 8 channels, stridding 2 × 2 × 2) as the main building block. The overall network stacks 4 of these main building blocks followed by two fully connected layers (16 and 2 units) with totally 0.3 million trainable parameters. Each layer is followed by a Rectified Linear Unit (ReLU) nonlinearity, except for the output layer where a Softmax activation is applied. Dimensionalities between layers describe the tensor size after each network layer.

Classifier network. We based our classifier on the 3D subject-level classifier network in[43]. Although the proposed network is reported to perform quite well, the number of trainable parameters (42 millions) relative to the dataset size is high, thus rendering it prone to overfitting. Hence, the number and size of the convolutional and fully connected layers were reduced until the network stopped overfitting on the training data and the validation accuracy started to drop. Batch normalization layers did not influence the performance of the network and were therefore removed. Finally, we replaced the max pooling layers by convolutional layers with striding as tested in[32]. Avoiding max pooling layers improves the interpretability of networks[44]. Dropout was not applied in the network. All biases in the classifier were constrained to be negative or zero to contribute to sparsify the network activations, and therefore, also to improve interpretability[44]. The final 3D classifier network is combining a single convolutional layer (kernel 3 × 3 × 3, 8 channels) with a down-convolutional layer (kernel 3 × 3 × 3, 8 channels, stridding 2 × 2 × 2) as the main building block. The overall network stacks 4 of these main building blocks followed by two fully connected layers (16 and 2 units) with totally 0.3 million trainable parameters. Each layer is followed by a Rectified Linear Unit (ReLU) nonlinearity, except for the output layer where a Softmax activation is applied. A graphical description of the classifier network is given in Fig. 1.

Heat mapping. Heat maps were created based on the deep Taylor decomposition (DTD) method described in[35]. This method is equivalent to the layer-wise relevance propagation rule LRP-$\alpha_1\beta_0$ for networks like the one we used in this study. The principal idea of DTD is to compute a Taylor decomposition of the relevance at a given network layer onto the lower layer. The name "deep Taylor decomposition" comes from the iterative application of Taylor decomposition from the top layer down to the input layer[44]. The output of the Softmax layer of the classifier network defines the relevance that is redistributed with this saliency method. With DTD the relevance is routed only along the positively contributing parts of the network. This is a desired property because we want to focus the network on brain regions with features that cause the classification result. Nevertheless, it is of importance to select a heat mapping method that passes simple sanity checks and is dependent on the network and the training[45,46]. While lower layers could become less influencing on the saliency map[47], DTD was shown to pass the sanity checks by computing saliency maps for all classes and then removing less relevant pixels from the final map[48]. Due to the nature of brain MRI data, we extended the currently available implementation of DTD from[49] to full 3D. The heat mapping method is used for both the relevance-guided classifier network and visualization.

Relevance-guided classifier network. The proposed relevance-guided network architecture focuses the classifier network on *relevant features* by extending the given network (cf. Fig. 2 top) with a relevance map generator (cf. Fig. 2 bottom). To this end we implemented the deep Taylor decomposition (z^+-rule) to generate the relevance maps of each input image depending on the classifier's current parameters during training, yielding the Graz$^+$ technique (guided relevance by adaptive z^+-rule).

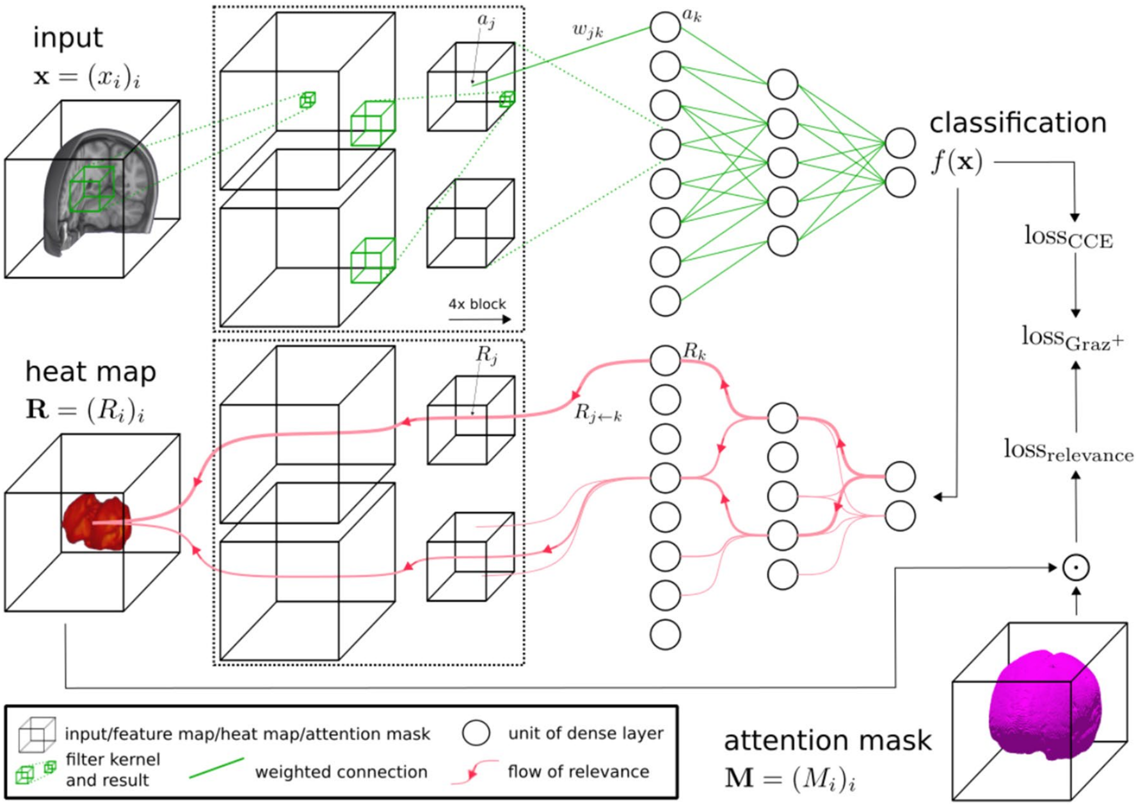

Figure 2. Schematic overview of the Graz$^+$ network and the adapted training process. A conventional classifier network (top) is extended by the heat map generator (bottom). For each classifier network layer a corresponding relevance redistribution layer with shared parameters and activations is attached to the generator network. The online calculated heat map is guiding the classifier training by adding a relevance sum inside the binary attention mask ($\mathrm{loss}_{\mathrm{relevance}}$), which is added to the categorical cross entropy loss ($\mathrm{loss}_{\mathrm{CCE}}$), yielding the total loss ($\mathrm{loss}_{\mathrm{Graz^+}}$). $\odot$ denotes the element-wise product. The solid green lines in the classifier network are the weighted connections between units of the layers, whereas items connected with dashed green lines symbolize information flow from one layer to the following layer (e.g. input to feature map, multiple feature maps to feature map or feature maps to densely connected unit). In contrast, the red lines in the heat map generator indicate the amount of relevance (thickness of line) flowing through the units of the layers to the heat map. If a unit does not contribute to the heat map the flow is stopped.

Loss function. In order to guide the training process by the attention mask (**M**), we extended the classifier's categorical cross entropy loss ($\mathrm{loss}_{\mathrm{CCE}}$) by a relevance-guided loss term to act as a regularizer:

$$\mathrm{loss}_{\mathrm{relevance}}(\mathbf{R}, \mathbf{M}) = -\mathbf{1}^T \mathrm{vec}(\mathbf{R} \odot \mathbf{M}), \tag{1}$$

which enforces higher relevance values in **R** at positions marked by the focus region **M**. Because the sum of relevance values in **R** is constant[35] this regularizer also automatically decreases the relevance values at positions not marked by **M**. Consequently, this approach yields the total loss per data sample:

$$\begin{aligned} \mathrm{loss}_{\mathrm{Graz^+}} &= \mathrm{loss}_{\mathrm{relevance}} + \mathrm{loss}_{\mathrm{CCE}} \\ &= -\mathbf{1}^T \mathrm{vec}(\mathbf{R} \odot \mathbf{M}) - \sum_{i=1}^{\mathrm{outputs}} y_i \cdot \log(\hat{y}_i), \end{aligned} \tag{2}$$

where **R** denotes the relevance heat map (3D shape), **M** is the predefined binary attention mask obtained during image preprocessing (3D shape), vec(**A**) denotes the row major vector representation of **A** resulting in a column vector (1D shape), and **1** is a column vector where all elements are set to 1 (1D shape). The inner product of the transposed vector **1** and the vector representation of $\mathbf{R} \odot \mathbf{M}$ gives the scalar value $\mathrm{loss}_{\mathrm{relevance}}$ (0D shape). The negative sign accounts for the maximization of the relevance inside the mask and $\odot$ denotes the element-wise product. For the categorical cross entropy y_i is the target value of the i-th output class and $\hat{y}_i$ its predicted value.

Hyperparameter optimization and training. Hyperparameter search on learning rate and learning rate schedule was done as proposed in CIFAR10-VGG11 experiments, in detail described in[50]. The batch size was omitted for consistent memory usage and exponential decay applied for learning rate schedule. Briefly, the hyperparameter optimization resulted in using the Adam optimizer with learning rate set to 10^{-4}, γ set to 1.0, β_1

Classifier	Skull stripping	Registration	Balanced accuracy	Sensitivity	Specificity	AUC
CNN	No	–	71.26 ± 2.86%	55.55 ± 7.51%	86.96 ± 3.95%	0.75 ± 0.02
		Lin.	74.27 ± 3.83%	63.13 ± 9.05%	85.40 ± 6.45%	0.80 ± 0.05
		Nonlin.	77.61 ± 4.44%	64.79 ± 5.02%	90.43 ± 5.19%	0.85 ± 0.06
CNN	Yes	–	77.66 ± 4.39%	69.70 ± 7.65%	85.63 ± 4.06%	0.83 ± 0.05
		Lin.	79.45 ± 3.34%	76.87 ± 4.81%	82.03 ± 6.23%	0.86 ± 0.05
		Nonlin.	82.13 ± 5.08%	73.47 ± 7.89%	90.78 ± 4.92%	0.88 ± 0.05
CNN+Graz$^+$	No	–	80.66 ± 4.80%	74.95 ± 7.85%	86.36 ± 2.85%	0.88 ± 0.04
		Lin.	**86.19 ± 6.01%**	79.73 ± 10.72%	**92.66 ± 3.73%**	**0.92 ± 0.04**
		Nonlin.	83.50 ± 5.90%	77.16 ± 8.95%	89.83 ± 4.49%	0.90 ± 0.04
Logistic regression*	Yes	Lin.**	82.00 ± 4.25%	**80.57 ± 7.16%**	83.43 ± 2.45%	0.90 ± 0.04

Table 1. Mean performance (in %) for the different models on all holdout data sets of cross validation. Highest values per column are highlighted in bold. *AUC* area under the curve of the receiver operating characteristics. *Logistic regression by FSL-SIENAX. **Linear registration is applied during FSL-SIENAX processing to obtain scaling factor.

set to 0.9, β_2 set to 0.999 and ε set to 10^{-7}[51] for 60 epochs with a batch size of 8 for training in all configurations. Each model was end-to-end trained with standard loss minimization and error backpropagation. We trained models for 3 differently preprocessed T1-weighted input images

- in native subject space,
- linearly registered to MNI152 template and
- nonlinearly registered to MNI152 template

and all cases were tested in

- standard classifier network with native images,
- standard classifier network with the skull removed and
- our relevance-guided method with predefined attention masks,

creating overall nine models. No data augmentation was used.

Cross validation.

AD and HC data were split up randomly into five folds without a separate test set, while maintaining all scans from one person in the same fold[43]. Final folds were created by combining one fold from each cohort to ensure class distribution within. The difference in the class sizes was accounted for using a class weighting in the loss function.

Model selection.

The optimal models based on the standard classifier networks were selected by highest validation classification accuracy. The relevance inside the attention mask threshold was set to 90% for the Graz$^+$ networks, enforcing models where most of the relevance is inside the intracranial volume.

Relevance-weighted heat map representation.

Besides qualitatively investigating individual heat maps, we calculated mean heat maps and histogram for each mean heatmap. Starting with the bin with the highest relevance values, the bin contents were added up until 50% of all relevance was included. The lower value of the last bin added was used as the lower value for windowing the mean heatmap. All heat maps shown in this paper are overlaid on the MNI152 1 mm template and windowed to present the top 50% of relevance.

Relevance density.

The relevance density describes the contribution of individual voxels of the heat map to the classification result. For all models we compare how many voxels are necessary to reach a certain level of explanation, e.g. how many voxels are needed to explain 85% of the total relevance.

Volumetry.

For comparison between deep learning and logistic regression models for AD classification, we calculated whole brain, gray matter as well as ventricular volume using FSL-SIENAX with a fractional intensity threshold of 0.35 and bias field/neck cleanup enabled[52].

Results

Model performances.

Table 1 reports the mean performance for the cross validation setup of all tested configurations. In summary:

- While models with skull stripping perform better than those without, the Graz$^+$ models yield even better balanced accuracy.

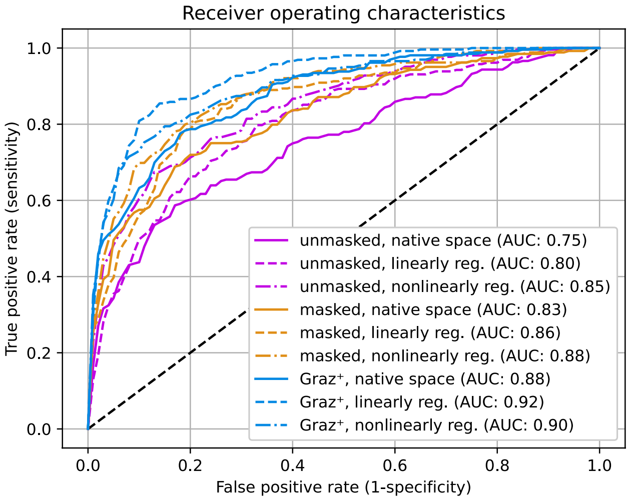

Figure 3. Comparison of mean receiver operating characteristics curves for all nine configurations. The Graz$^+$ models (blue) show higher values for the area under the curve (AUC in legend) compared to unmasked (purple) and masked (orange) configurations.

- The Graz$^+$ model with linearly registered input had the highest balanced accuracy (86.19%), AUC (0.92) and also specificity (92.66%).
- Linear and nonlinear registration improves the balanced accuracy independently of skull stripping and utilization of Graz$^+$.
- The logistic regression model based on volumetric information for the entire brain, gray matter, and ventricular volume yielded a balanced accuracy of 82.00%, which is comparable or even outperforming some CNN models without skull stripping.

As the used dataset is nearly balanced[53], the corresponding mean receiver operating characteristics (ROC) curves for these models are shown in Fig. 3.

Heat mapping. Mean heat maps for classification decisions on cross validation holdout data sets for all trained models are shown in Fig. 4, overlaid on the MNI152 1 mm template. Individual heatmaps were nonlinearly transformed to the MNI152 space before averaging. Transformation information was obtained during T1-weighted image preprocessing. Visual inspection of the heat maps reveals that the processing type (unmasked/masked/Graz$^+$) yields substantially different results (columns), while the impact of the registration type (no registration/linear/nonlinear) is rather limited. Although mean heat maps in each column appear visually similar, applying registration to input MRIs improves the balanced accuracy. When using the native T1-weighted images as input, the most relevant features are obtained in the scalp/skull outside brain parenchyma (unmasked configurations, left column). When skull stripping of the input MRIs is applied, the highest relevances are found in the cerebral and cerebellar cortex or generally adjacent to the brain-CSF-interface (middle column). While the aforementioned classifiers also show minor relevances in central brain regions, the maps from Graz$^+$ show relevant regions exclusively within deep gray and white matter tissue adjacent to the ventricles (right column). Figure 5 shows multiple slices of mean heat maps for classification decisions of all cross validation holdout data sets for all trained models, overlaid on the MNI152 template.

Relevance density. Figure 6 shows that the Graz$^+$ training increased the sparsity of the utilized features, where the 10% most relevant voxels (x-axis) explain approximately 20% (unmasked), 35% (masked) and 75% (Graz$^+$) of the total relevance.

Discussion

Summary. The present work investigated the mechanisms underlying brain disease classification by CNNs. Understanding the classifier's decision(s) is highly relevant, not only from an ethno-clinical but particularly from a legal perspective. We demonstrated how dramatic T1-weighted Alzheimer's disease classification is depending on volumetric features. Moreover, we show that preprocessing of neuroimaging data is decisive for feature identification because it introduces novel misleading features subsequently utilized for classification. The presented Graz$^+$ technique is addressing these issues by focusing the feature identification on the intracranial space only. This yields higher classification accuracy than conventional CNN-methods, but more importantly, it substantially resolves the impact of MR image preprocessing.

Impact to deep learning-based neuroimaging studies. Our motivation for this work was driven by simple recurring questions in clinical brain MRI studies: Should the skull from a conventional T1-weighted

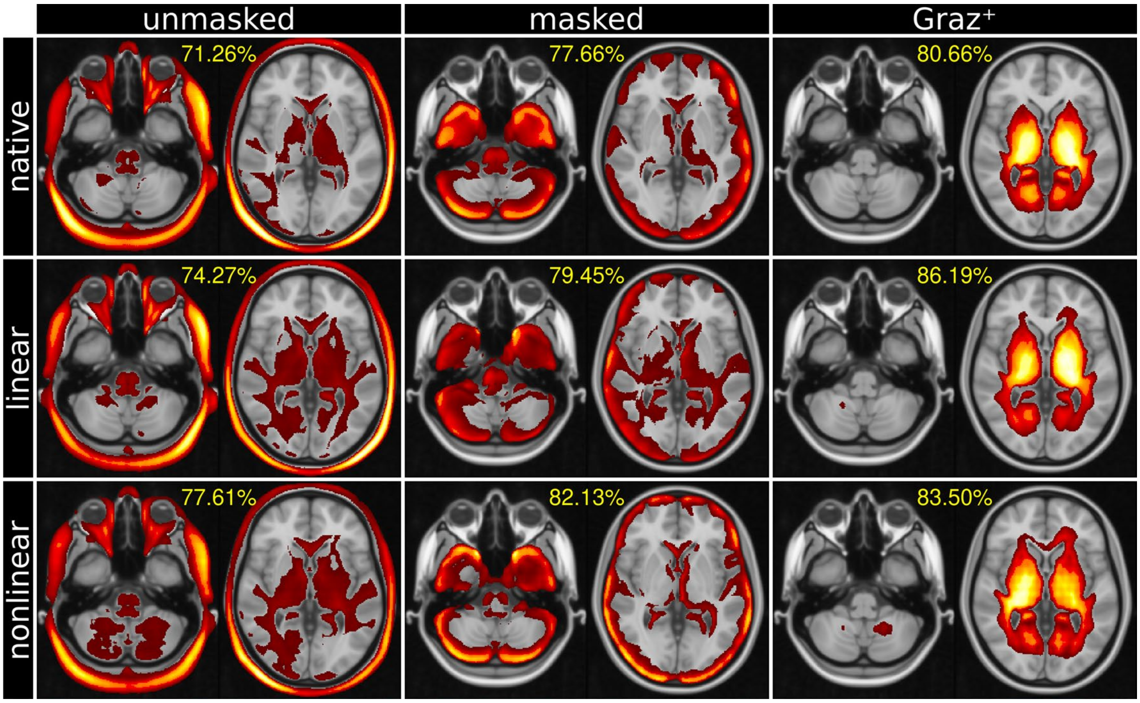

Figure 4. Mean heat maps (highest relevance in yellow, overlaid on MNI152 template) and balanced classification accuracy (percentage). Unmasked and masked CNN classifiers obtain relevant image features overwhelmingly from global volumetric information (left and center columns), whereas Graz$^+$ exclusively relies on deep gray and white matter tissue adjacent to the ventricles (right column). Heat maps are thresholded to the top 50% of the overall relevance. See "Methods" for description.

MRI be stripped for further processing or should the entire MRI including skull and neck be used? Additionally, whether and which type of image registration is required or best as the next preprocessing step?

Showing that the preprocessing of MR images is crucial for the feature identification by CNNs has severe implications for neuroimaging based machine learning classifications. A majority of analysis pipelines apply skull stripping during image processing. This avoids the identification of features outside of the brain tissue, but in turn introduces new edges at the newly created brain mask, which might be subsequently used by the CNN for classification. We anticipate that decisions also might be misled by underlying contributors such as the implementation of the skull stripping algorithm, brain atrophy, but also might reflect visually not observable information as involuntary patients' movements. Generally, the source and extent of the newly introduced features remains unclear, however it was demonstrated that skull stripping algorithms can be biased by the patient cohort[54], thus, additionally biasing the classification. In Table 1 CNNs showed improved performance when skull stripping is applied and independent of registration. However, those models overwhelmingly highlighted regions at the brain-CSF-interface as shown in Fig. 5.

Addressing these shortcomings, the proposed relevance-guided Graz$^+$ method identified regions of highest relevance in brain parenchyma while the balanced accuracy remained comparable or even better. Moreover, pooling data from rare diseases or generally small datasets often yield potentially spurious results and low replicability[55]. Its invariance from registration and skull stripping methods provides a usable method for CNN-based classification studies which might be practically useful when pooling data from different scanners and sites[56] or assisting statistical harmonization[57,58].

While the Graz$^+$ method can be natively applied to multi center data, a better way would be to use bias field correction as described in[43] and additionally normalizing T1-weighted images to a specified intensity range. The present study included images from a single scanner using an identical MRI protocol for all subjects. Consequently, this reduces the need for the above mentioned steps and allowed to focus on the effects of brain masking and registration.

Neuroanatomical and biophysical interpretation.

This section highlights plausible mechanisms underlying CNN-based disease classification in AD by analyzing the neuroanatomical position of voxel relevance observed by heat mapping. The highest relevances were observed in the scalp for the CNN models using native (unmasked) input images. With skull stripping (masked), the most relevant voxels were found at the brain-CSF-interface, respectively, at the newly-introduced edges of the brain parenchyma. Anatomically, these regions are substantially overlapping with cortical gray matter, where atrophy is a well-known effect in AD. Cortical gray matter changes might be reflected in the masked CNNs decision, but seem rather implausible because of the small magnitude compared to global atrophy and ventricular enlargement. However, we cannot entirely rule out a secondary effect from the brain extraction algorithm biased by the patient cohort[54]. Both CNN meth-

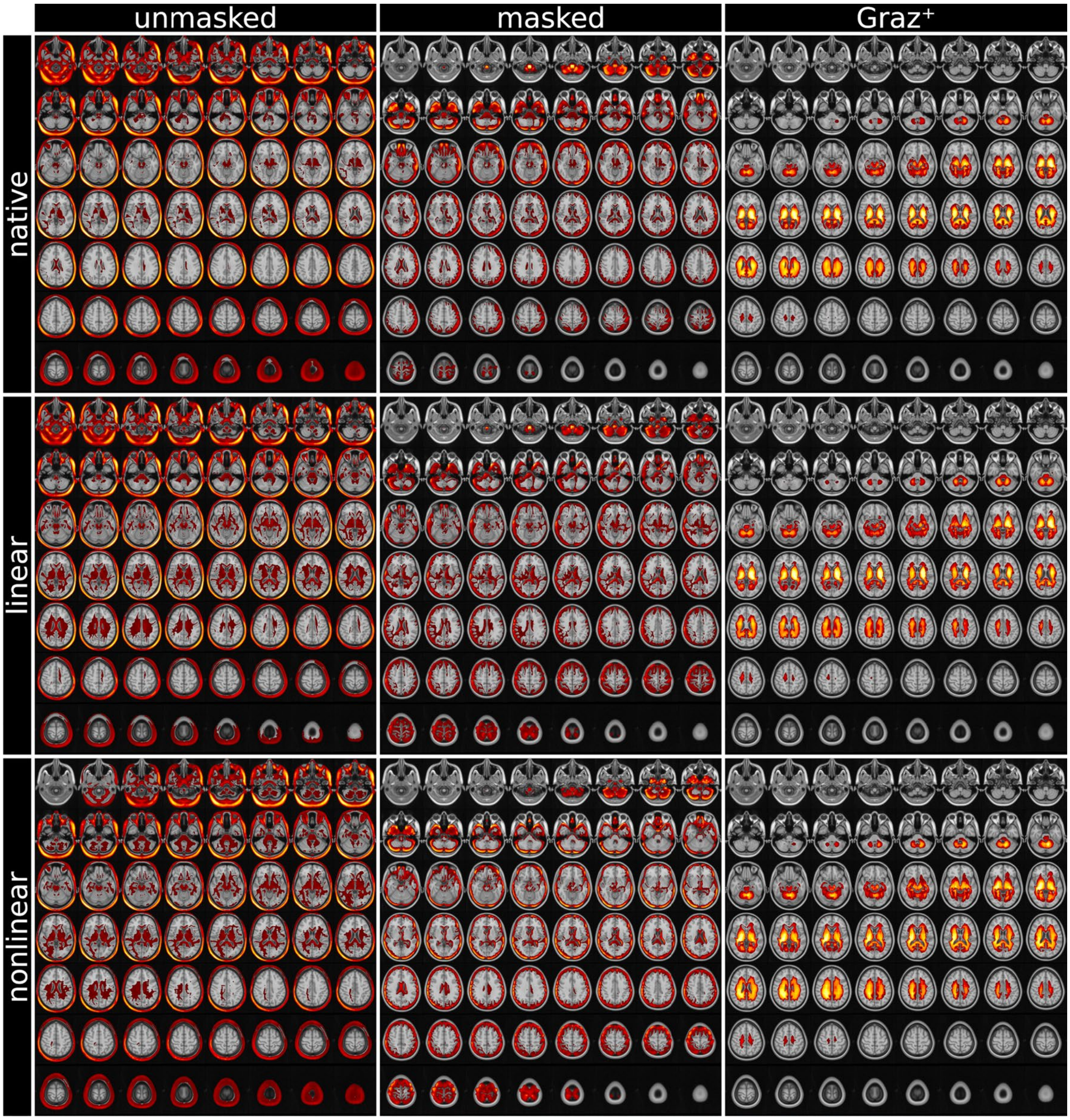

Figure 5. MNI152 template overlaid by mean relevance maps (highest relevance in yellow) obtained for all nine models. Unmasked and masked MRI classifiers obtain relevant image features from volumetric information (left and center columns). In contrast, the proposed Graz$^+$-method bases the classifier's decision on deep brain image features, virtually independently of the registration method (right column). Heat maps are thresholded to the top 50% of the overall relevance. See "Methods" for description.

ods also identified some relevant voxel clusters in deep gray and white matter adjacent to the lateral ventricles (center of the brain), which were substantially smaller.

Given the spatial distribution of the relevances, we argue that the two conventional CNN models are overwhelmingly sensitive for global volumetric features. Complementary volumetric analysis using an established neuroimaging tool for brain segmentation (FSL-SIENAX) in a logistic regression model yielded a balanced accuracy of 82%, which is on par with the top CNN results. Here we want to highlight that FSL-SIENAX volumetry did not rely on the brain masks utilized by the CNNs, but instead uses inversely transformed brain masks of the MNI152 brain and FSL-FAST to estimate the volumes.

The Graz$^+$-based models identified regions with highest relevance mainly in deep gray and white matter located adjacent to the lateral ventricles. However, the anatomical/biophysical underpinnings of the decisions are less clear than in the conventional CNN models. Beside aforementioned contributions of volumetric features (AD progression is commonly paralleled by ventricular enlargement and global atrophy) also the T1-weighted contrast can pathologically change in AD[59]. White matter hyperintensities (WMH) are commonly seen in brain MRI in older people and beside their underlying heterogeneous histopathology, they represent radiological correlates of cognitive and functional impairment[60]. In a previous study, we found WMHs preferentially in a bilateral periventricular location, partly overlapping with the regions identified here by the Graz$^+$-based models[61]. Furthermore,

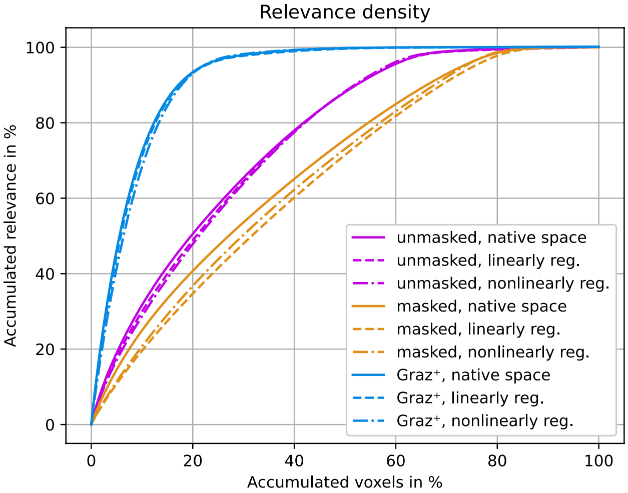

Figure 6. The relevance density describes the contribution of individual voxels to the classification decision. Removal of scalp tissue voxels (orange) yields higher relevance density compared to unmasked T1 images (purple). The Graz$^+$-models (blue) identify sparser but substantially more relevant voxels, which improves the classification accuracy.

other plausible contributors are increased brain iron deposition in the deep gray matter (basal ganglia) of AD patients[40] or cumulative gadolinium deposition of macrocyclic contrast agents[62]. Nevertheless, with the given setup we cannot definitely disentangle the underlying constituents and refer to the validation section below.

The relevance density analysis revealed that Graz$^+$-based models learn much sparser features, subsequently needing less voxels for inferring classification decisions. Consequently, we hypothesize that the lack of misleading voxels from the scalp or newly-introduced edges is responsible for the increased accuracy.

Related work.

With the availability of accessible large MRI databases from patients, such as the Alzheimer's Disease Neuroimaging Initiative (ADNI), AIBL or OASIS databases, various studies using machine learning techniques exploiting structural imaging data have been published, formerly using *classical* machine learning classification methods (e.g. LDA, SVM) in combination with feature extraction methods based on tissue density[63], cortical surface[64] and hippocampal measurements[65]. Reported classification accuracies range between 75% and 100%, comprehensively summarized in[66]. Recently, interests switched to deep learning CNNs for (A) classification[14,15,17], (B) classification with explanation[19,20,67] and (C) regression with explanation[18] of AD. A recent review summarizes the state-of-the-art using CNNs for AD classification, comparing various network architectures, input data and disease subtypes[43]. Strictly in line with the data leakage analysis in this work we utilized stratified cross validation, while maintaining all datasets from one person in the same fold. Furthermore, we used the input MR images in their native spatial resolution, avoiding unpredictable influence from down- or resampling. While most of the analyzed studies are based on the ADNI dataset, our classification performance results are on par with both remaining 3D subject-level approaches without data leakage[14,21].

The inconsistency between learned features with linear and nonlinear registration is systematically investigated in[18]. They found that the use of nonlinearly registered images to train CNNs can drive the network by registration artifacts, which may also explain the higher performance for linearly registered images compared to nonlinearly registered images as input for the CNN+Graz$^+$ in Table 1. However, the influence of further preprocessing steps on the resulting models and performances is less well known. Heat mapping using the LRP framework has been sparsely applied for explaining the underpinnings of an AD diagnosis in convolutional neural networks trained with structural MRI data beside the work of[20]. Heat maps obtained by two techniques (LRP and guided backpropagation) indicate relevant features in the parahippocampal gyrus but also adjacent to the brain-CSF interface, which is in line with our work.

Regularized heat map learning has been proposed before, however, differently to the Graz$^+$ method integrating a-priori knowledge with predefined attention masks. Technically, the gradient of the function learned by the network with respect to the current input can be interpreted as a heat map[31]. Regularization of this input gradient was first introduced by[68] as *double back-propagation*, which trains neural networks by not only minimizing the *energy* of the network but the rate of change of that energy with respect to the input features. In[69] this regularization was extended by selectively penalizing the gradient. Whereas[70] use LRP to create maps during training, which are multiplied with the corresponding input and then fed to the original classifier to dynamically find and emphasize important features. Furthermore, attention gated networks for medical image analysis have been proposed to automatically learn to focus on target structures of varying shapes and sizes[71].

Validation.

Direct validation of the classifier's decision is generally hardly feasible in the absence of a ground truth. While we anticipate a correspondence of the volumetric features with Alzheimer's atrophy, this conclusion might not be final. However, in future work, indirect validation is possible using quantitative MRI parameters such as relaxometry, susceptibility, or magnetization transfer, where regional effects are known from ROI-based, voxel-based morphometry (VBM) or radiomics studies. While those methods statistically assess neuroanatomical features including ventricular enlargement or hippocampal atrophy, quantitative MRI parameters describe the underlying biophysical tissue composition. The effective relaxation rate R_2^* can assess increased iron deposition in the basal ganglia, a frequent finding in AD[40]. Consequently, the potential overlap with heat maps in those regions is better suited to disentangle biophysical tissue changes from atrophy. Optionally, direct validation of our method would require the generation of a cohort of realistic in silico phantoms (as recently used in the quantitative susceptibility mapping (QSM) image reconstruction challenge 2.0[72]) with modulateable regional relaxation times in conjunction with an adjustable atrophy deformator[73,74].

Limitations. Several aforementioned neuroimaging studies used the ADNI (or other publicly available) database for deep learning based classification. Generally, the clinical relevance of an automated AD classification is limited. The prodromal state of mild cognitive impairment (MCI) is preceding AD and identification of individuals rapidly progressing to AD (or differential diagnosis of frontotemporal dementia types) would be of higher importance for clinical management. We acknowledge the absence of an MCI group as a limitation and therefore provide the source code for the fast reproducibility using alternative network topologies, input data (quantitative MRI, PET), and other diseases. While aforementioned databases are designed multi-centrically, all MRI scans used in this paper were acquired with a single 3T scanner. Beside the underlying AD patient data, comparison with other studies is hampered by different network architectures, preprocessing and hyperparameter selection[43].

While this study only applied whole brain masks, more focused masks guiding the attention to e.g. the precuneus, the entorhinal cortex, the parietal lobe, the temporal lobe or the hippocampi are feasible, especially when regional a-priori knowledge for a certain pathology exists. Because of the explorative nature of the novel methodological framework we focused on the brain parenchyma. Organs outside the brain are more variable in size and shape, which render registration and ROI-definition more challenging. We originally developed Graz$^+$ for clinical brain studies, but its invariance to preprocessing might be even more pronounced beyond neuroimaging.

Lastly, the absence of CSF biomarkers or amyloid/tau-PET for the AD diagnosis reduces the accuracy of the clinical diagnosis. However, AD diagnosis using the NINCDS-ADRDA criteria has a sensitivity of 81% and specificity of 70% as shown in clinico-pathological studies[39].

Conclusion

This work highlights that CNNs are not necessarily more efficient or better regarding classification accuracy than simple conventional volumetric features. However, the proposed relevance-guided approach is neutralizing the impact of MRI preprocessing from skull stripping and registration, rendering it a practically usable and robust method for CNN-based neuroimaging classification studies. Relevance-guiding focuses feature identification on the intracranial space only, yielding physiological plausible results and as a secondary effect the classification accuracy is higher (Supplementary information S1).

Data availability

Source code for Graz$^+$ and the image preprocessing is available under www.neuroimaging.at/explainable-ai. The MR images used in this paper are part of a clinical data set and therefore are not publicly available. Formal data sharing requests to the corresponding author will be considered.

Received: 25 April 2022; Accepted: 16 November 2022
Published online: 24 November 2022

References

1. Scheltens, P. *et al.* Alzheimer's disease. *Lancet (London, England)* **397**, 1577–1590. https://doi.org/10.1016/S0140-6736(20)32205-4 (2021).
2. Braak, H., Alafuzoff, I., Arzberger, T., Kretzschmar, H. & Del Tredici, K. Staging of Alzheimer disease-associated neurofibrillary pathology using paraffin sections and immunocytochemistry. *Acta Neuropathol.* **112**, 389–404. https://doi.org/10.1007/s00401-006-0127-z (2006).
3. Braak, H. & Braak, E. Neuropathological stageing of Alzheimer-related changes. *Acta Neuropathol.* **82**, 239–259. https://doi.org/10.1007/BF00308809 (1991).
4. Oldan, J. D., Jewells, V. L., Pieper, B. & Wong, T. Z. Complete Evaluation of Dementia: PET and MRI Correlation and Diagnosis for the Neuroradiologist. *AJNR Am. J. Neuroradiol.*https://doi.org/10.3174/ajnr.A7079 *(2021).*
5. Dubois, B. *et al.* Clinical diagnosis of Alzheimer's disease: recommendations of the International Working Group. *Lancet Neurol.* **20**, 484–496. https://doi.org/10.1016/S1474-4422(21)00066-1 (2021).
6. Biel, D. *et al.* Tau-PET and in vivo Braak-staging as a prognostic marker in Alzheimer's disease. *medRxiv* 2021.02.04.21250760. https://doi.org/10.1101/2021.02.04.21250760 (2021).
7. Vogel, J. W. *et al.* Four distinct trajectories of tau deposition identified in Alzheimer's disease. *Nat. Med.*https://doi.org/10.1038/s41591-021-01309-6 *(2021).*
8. Henneman, W. J. P. *et al.* Hippocampal atrophy rates in Alzheimer disease: Added value over whole brain volume measures. *Neurology* **72**, 999–1007. https://doi.org/10.1212/01.wnl.0000344568.09360.31 (2009).
9. Leung, K. K. *et al.* Cerebral atrophy in mild cognitive impairment and Alzheimer disease: Rates and acceleration. *Neurology* **80**, 648–654. https://doi.org/10.1212/WNL.0b013e318281ccd3 (2013).
10. Sluimer, J. D. *et al.* Whole-brain atrophy rate in Alzheimer disease: Identifying fast progressors. *Neurology* **70**, 1836–1841. https://doi.org/10.1212/01.wnl.0000311446.61861.e3 (2008).
11. Hammernik, K. *et al.* Learning a variational network for reconstruction of accelerated MRI data. *Magn. Reson. Med.* **79**, 3055–3071. https://doi.org/10.1002/mrm.26977 (2018).
12. Kleesiek, J. *et al.* Deep MRI brain extraction: A 3D convolutional neural network for skull stripping. *Neuroimage* **129**, 460–469. https://doi.org/10.1016/j.neuroimage.2016.01.024 (2016).

13. Esteva, A. *et al.* Dermatologist-level classification of skin cancer with deep neural networks. *Nature* **542**, 115–118. https://doi.org/10.1038/nature21056 (2017).
14. Bäckström, K., Nazari, M., Gu, I. Y.-H. & Jakola, A. S. An efficient 3D deep convolutional network for Alzheimer's disease diagnosis using MR images. In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, 149–153. https://doi.org/10.1109/ISBI.2018.8363543 (2018). ISSN: 1945-8452.
15. Noor, M. B. T., Zenia, N. Z., Kaiser, M. S., Mamun, S. A. & Mahmud, M. Application of deep learning in detecting neurological disorders from magnetic resonance images: a survey on the detection of Alzheimer's disease: Parkinson's disease and schizophrenia. *Brain Inf.* **7**, 11. https://doi.org/10.1186/s40708-020-00112-2 (2020).
16. Vieira, S., Pinaya, W. H. L. & Mechelli, A. Using deep learning to investigate the neuroimaging correlates of psychiatric and neurological disorders: Methods and applications. *Neurosci. Biobehav. Rev.* **74**, 58–75. https://doi.org/10.1016/j.neubiorev.2017.01.002 (2017).
17. Zhang, L., Wang, M., Liu, M. & Zhang, D. A Survey on Deep Learning for Neuroimaging-Based Brain Disorder Analysis. *Front. Neurosci.* **14**. https://doi.org/10.3389/fnins.2020.00779 (2020).
18. Dinsdale, N. K. *et al.* Learning patterns of the ageing brain in MRI using deep convolutional networks. *Neuroimage* **224**, 117401. https://doi.org/10.1016/j.neuroimage.2020.117401 (2021).
19. Oh, K., Chung, Y.-C., Kim, K. W., Kim, W.-S. & Oh, I.-S. Classification and visualization of Alzheimer's disease using volumetric convolutional neural network and transfer learning. *Sci. Rep.* **9**, 18150. https://doi.org/10.1038/s41598-019-54548-6 (2019).
20. Böhle, M., Eitel, 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.* **11**, 194. https://doi.org/10.3389/fnagi.2019.00194 (2019).
21. Korolev, S., Safiullin, A., Belyaev, M. & Dodonova, Y. Residual and plain convolutional neural networks for 3D brain MRI classification. In *2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017)*, 835–838, https://doi.org/10.1109/ISBI.2017.7950647 (2017). ISSN: 1945-8452.
22. Karapinar Senturk, Z. Early diagnosis of Parkinson's disease using machine learning algorithms. *Med. Hypotheses* **138**, 109603. https://doi.org/10.1016/j.mehy.2020.109603 (2020).
23. Eitel, F. *et al.* Uncovering convolutional neural network decisions for diagnosing multiple sclerosis on conventional MRI using layer-wise relevance propagation. *NeuroImage. Clin.* **24**, 102003. https://doi.org/10.1016/j.nicl.2019.102003 (2019).
24. Davatzikos, C. Machine learning in neuroimaging: Progress and challenges. *Neuroimage* **197**, 652–656. https://doi.org/10.1016/j.neuroimage.2018.10.003 (2019).
25. Lapuschkin, S. *et al.* Unmasking Clever Hans predictors and assessing what machines really learn. *Nat. Commun.* **10**, 1096. https://doi.org/10.1038/s41467-019-08987-4 (2019).
26. Goodman, B. & Flaxman, S. European union regulations on algorithmic decision-making and a "right to explanation". *AI Mag.* **38**, 50–57. https://doi.org/10.1609/aimag.v38i3.2741 (2017).
27. OECD. *Artificial Intelligence in Society* (OECD, 2019).
28. Lapuschkin, S., Binder, A., Montavon, G., Müller, K. & Samek, W. Analyzing Classifiers: Fisher Vectors and Deep Neural Networks. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2912–2920. https://doi.org/10.1109/CVPR.2016.318 (2016). ISSN: 1063-6919.
29. Tjoa, E. & Guan, C. A Survey on Explainable Artificial Intelligence (XAI): Toward Medical XAI. *IEEE transactions on neural networks and learning systems***PP**. https://doi.org/10.1109/TNNLS.2020.3027314 (2020).
30. Ribeiro, M. T., Singh, S. & Guestrin, C. "Why Should I Trust You?" Explaining the Predictions of Any Classifier. arXiv:1602.04938 [cs, stat] (2016).
31. Simonyan, K., Vedaldi, A. & Zisserman, A. Deep inside convolutional networks: Visualising image classification models and saliency maps. *ICLR* (2014).
32. Springenberg, J. T., Dosovitskiy, A., Brox, T. & Riedmiller, M. Striving for Simplicity: The all convolutional net. arXiv:1412.6806 [cs] (2015).
33. Zeiler, M. D. & Fergus, R. Visualizing and Understanding Convolutional Networks. In Fleet, D., Pajdla, T., Schiele, B. & Tuytelaars, T. (eds.) *Computer Vision - ECCV 2014*, Lecture Notes in Computer Science, 818–833. https://doi.org/10.1007/978-3-319-10590-1_53 (Springer International Publishing, Cham, 2014).
34. Zintgraf, L. M., Cohen, T. S., Adel, T. & Welling, M. Visualizing deep neural network decisions: Prediction difference analysis. arXiv:1702.04595 [cs] (2017).
35. Montavon, G., Lapuschkin, S., Binder, A., Samek, W. & Müller, K.-R. Explaining nonlinear classification decisions with deep Taylor decomposition. *Pattern Recogn.* **65**, 211–222. https://doi.org/10.1016/j.patcog.2016.11.008 (2017).
36. Bach, S. *et al.* On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLoS ONE* **10**, e0130140. https://doi.org/10.1371/journal.pone.0130140 (2015).
37. Montavon, G. Gradient-Based Vs. Propagation-Based Explanations: An Axiomatic Comparison. In Samek, W., Montavon, G., Vedaldi, A., Hansen, L. K. & Müller, K.-R. (eds.) *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, Lecture Notes in Computer Science, 253–265. https://doi.org/10.1007/978-3-030-28954-6_13 (Springer International Publishing, Cham, 2019).
38. Samek, W., Binder, A., Montavon, G., Lapuschkin, S. & Müller, K. Evaluating the Visualization of What a Deep Neural Network Has Learned. *IEEE Transactions on Neural Networks and Learning Systems***28**, 2660–2673, https://doi.org/10.1109/TNNLS.2016.2599820 (2017). Conference Name: IEEE Transactions on Neural Networks and Learning Systems.
39. Knopman, D. S. *et al.* Practice parameter: diagnosis of dementia (an evidence-based review). Report of the Quality Standards Subcommittee of the American Academy of Neurology. *Neurology* **56**, 1143–1153. https://doi.org/10.1212/wnl.56.9.1143 (2001).
40. Damulina, A. *et al.* Cross-sectional and Longitudinal Assessment of Brain Iron Level in Alzheimer Disease Using 3-T MRI. *Radiology* **296**, 619–626. https://doi.org/10.1148/radiol.2020192541 (2020) (**Publisher: Radiological Society of North America.**).
41. Schmidt, R. *et al.* Progression of cerebral white matter lesions: 6-year results of the Austrian Stroke Prevention Study. *Lancet (London, England)* **361**, 2046–2048. https://doi.org/10.1016/s0140-6736(03)13616-1 (2003).
42. Smith, S. M. *et al.* Advances in functional and structural MR image analysis and implementation as FSL. *Neuroimage* **23**(Suppl 1), S208-219. https://doi.org/10.1016/j.neuroimage.2004.07.051 (2004).
43. Wen, J. *et al.* Convolutional neural networks for classification of Alzheimer's disease: Overview and reproducible evaluation. *Med. Image Anal.* **63**, 101694. https://doi.org/10.1016/j.media.2020.101694 (2020).
44. Montavon, G., Samek, W. & Müller, K.-R. Methods for interpreting and understanding deep neural networks. *Digital Signal Process.* **73**, 1–15. https://doi.org/10.1016/j.dsp.2017.10.011 (2018).
45. Adebayo, J. *et al.* Sanity checks for saliency maps. arXiv:1810.03292 [cs, stat] (2020).
46. Yona, G. & Greenfeld, D. Revisiting Sanity Checks for Saliency Maps. arXiv:2110.14297 [cs] (2021).
47. 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*, 9046–9057 (PMLR, 2020). ISSN: 2640-3498.
48. Gupta, A. & Arora, S. A Simple Saliency Method That Passes the Sanity Checks. arXiv:1905.12152 [cs, stat] (2019).
49. Alber, M. *et al.* iNNvestigate Neural Networks!. *J. Mach. Learn. Res.* **20**, 1–8 (2019).
50. Bouthillier, X. *et al.* Accounting for Variance in Machine Learning Benchmarks. arXiv:2103.03098 [cs, stat] (2021).
51. Kingma, D. P. & Ba, J. Adam: A Method for Stochastic Optimization. *ICLR* (2015).

52. Smith, S. M. *et al.* Accurate, robust, and automated longitudinal and cross-sectional brain change analysis. *Neuroimage* **17**, 479–489. https://doi.org/10.1006/nimg.2002.1040 (2002).
53. Saito, T. & Rehmsmeier, M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS ONE* **10**, e0118432. https://doi.org/10.1371/journal.pone.0118432 (2015) (**Publisher: Public Library of Science.**).
54. Fennema-Notestine, C. *et al.* Quantitative evaluation of automated skull-stripping methods applied to contemporary and legacy images: Effects of diagnosis, bias correction, and slice location. *Hum. Brain Mapp.* **27**, 99–113. https://doi.org/10.1002/hbm.20161 (2006).
55. Varoquaux, G. Cross-validation failure: Small sample sizes lead to large error bars. *Neuroimage* **180**, 68–77. https://doi.org/10.1016/j.neuroimage.2017.06.061 (2018).
56. Clarke, W. T. *et al.* Multi-site harmonization of 7 tesla MRI neuroimaging protocols. *Neuroimage* **206**, 116335. https://doi.org/10.1016/j.neuroimage.2019.116335 (2020).
57. Dinsdale, N. K., Jenkinson, M. & Namburete, A. I. L. Deep learning-based unlearning of dataset bias for MRI harmonisation and confound removal. *Neuroimage* **228**, 117689. https://doi.org/10.1016/j.neuroimage.2020.117689 (2021).
58. Pomponio, R. *et al.* Harmonization of large MRI datasets for the analysis of brain imaging patterns throughout the lifespan. *Neuroimage* **208**, 116450. https://doi.org/10.1016/j.neuroimage.2019.116450 (2020).
59. Besson, J. a. O. *et al.* Nuclear Magnetic Resonance (NMR) II. Imaging in Dementia. *The British Journal of Psychiatry***146**, 31–35. https://doi.org/10.1192/bjp.146.1.31 (1985). Publisher: Cambridge University Press.
60. Prins, N. D. & Scheltens, P. White matter hyperintensities, cognitive impairment and dementia: an update. *Nature Reviews Neurology***11**, 157–165, https://doi.org/10.1038/nrneurol.2015.10 (2015). Bandiera_abtest: a Cg_type: Nature Research Journals Number: 3 Primary_atype: Reviews Publisher: Nature Publishing Group Subject_term: Alzheimer's disease;Brain imaging;Dementia Subject_term_id: alzheimers-disease;brain-imaging;dementia.
61. Damulina, A. *et al.* White matter hyperintensities in Alzheimer's disease: A lesion probability mapping study. *J. Alzheimer's Dis. JAD* **68**, 789–796. https://doi.org/10.3233/JAD-180982 (2019).
62. Kanda, T., Ishii, K., Kawaguchi, H., Kitajima, K. & Takenaka, D. High signal intensity in the dentate nucleus and globus pallidus on unenhanced T1-weighted MR images: relationship with increasing cumulative dose of a gadolinium-based contrast material. *Radiology* **270**, 834–841. https://doi.org/10.1148/radiol.13131669 (2014).
63. Klöppel, S. *et al.* Automatic classification of MR scans in Alzheimer's disease. *Brain J. Neurol.* **131**, 681–689. https://doi.org/10.1093/brain/awm319 (2008).
64. Eskildsen, S. F. *et al.* Prediction of Alzheimer's disease in subjects with mild cognitive impairment from the ADNI cohort using patterns of cortical thinning. *Neuroimage* **65**, 511–521. https://doi.org/10.1016/j.neuroimage.2012.09.058 (2013).
65. Sørensen, L. *et al.* Early detection of Alzheimer's disease using MRI hippocampal texture. *Hum. Brain Mapp.* **37**, 1148–1161. https://doi.org/10.1002/hbm.23091 (2016).
66. Rathore, S., Habes, M., Iftikhar, M. A., Shacklett, A. & Davatzikos, C. A review on neuroimaging-based classification studies and associated feature extraction methods for Alzheimer's disease and its prodromal stages. *Neuroimage* **155**, 530–548. https://doi.org/10.1016/j.neuroimage.2017.03.057 (2017).
67. Tang, Z. *et al.* Interpretable classification of Alzheimer's disease pathologies with a convolutional neural network pipeline. *Nat. Commun.* **10**, 2173. https://doi.org/10.1038/s41467-019-10212-1 (2019) (**Number: 1 Publisher: Nature Publishing Group.**).
68. Drucker, H. & Le Cun, Y. Improving generalization performance using double backpropagation. *IEEE Trans. Neural Netw.* **3**, 991–997. https://doi.org/10.1109/72.165600 (1992).
69. Ross, A. S., Hughes, M. C. & Doshi-Velez, F. Right for the Right Reasons: Training Differentiable Models by Constraining their Explanations. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*, 2662–2670, https://doi.org/10.24963/ijcai.2017/371 (International Joint Conferences on Artificial Intelligence Organization, Melbourne, Australia, 2017).
70. Sun, J. *et al.* Explanation-Guided Training for Cross-Domain Few-Shot Classification. In *2020 25th International Conference on Pattern Recognition (ICPR)*, 7609–7616. https://doi.org/10.1109/ICPR48806.2021.9412941 (2021). ISSN: 1051-4651.
71. Schlemper, J. *et al.* Attention gated networks: Learning to leverage salient regions in medical images. *Med. Image Anal.* **53**, 197–207. https://doi.org/10.1016/j.media.2019.01.012 (2019).
72. Marques, J. P. *et al.* QSM reconstruction challenge 2.0: A realistic in silico head phantom for MRI data simulation and evaluation of susceptibility mapping procedures. *Magn. Reson. Med.***86**, 526–542, https://doi.org/10.1002/mrm.28716 (2021). https://onlinelibrary.wiley.com/doi/pdf/10.1002/mrm.28716.
73. Khanal, B., Ayache, N. & Pennec, X. Simulating longitudinal brain MRIs with known volume changes and realistic variations in image intensity. *Front. Neurosci.* **11**. https://doi.org/10.3389/fnins.2017.00132 (2017).
74. Khanal, B., Lorenzi, M., Ayache, N. & Pennec, X. A biophysical model of brain deformation to simulate and analyze longitudinal MRIs of patients with Alzheimer's disease. *Neuroimage* **134**, 35–52. https://doi.org/10.1016/j.neuroimage.2016.03.061 (2016).

Acknowledgements

This study was funded by the Austrian Science Fund (FWF grant numbers: KLI523, P30134). This research was supported by NVIDIA GPU hardware grants.

Author contributions

C.T.: conceptualization, data curation, formal analysis, investigation, methodology, software, validation, visualization, writing - original draft preparation, S.H.: formal analysis, validation, writing - review and editing, L.P.: data curation, validation, writing - review and editing, A.D.: data curation, writing - review and editing, R.Sc.: data curation, writing - review and editing, R.St.: supervision, writing - review and editing, S.R.: data curation, validation, writing - review and editing, C.L.: conceptualization, data curation, formal analysis, funding acquisition, investigation, methodology, project administration, resources, supervision, validation, writing - original draft preparation.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-022-24541-7.

Correspondence and requests for materials should be addressed to C.L.

Reprints and permissions information is available at www.nature.com/reprints.

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