Contents lists available at ScienceDirect

Artificial Intelligence In Medicine

journal homepage: www.elsevier.com/locate/artmed

Research paper

Investigating the discrimination ability of 3D convolutional neural networks applied to altered brain MRI parametric maps

Giulia Maria Mattia [a,*], Edouard Villain [b,a], Federico Nemmi [a], Marie-Véronique Le Lann [b], Xavier Franceries [c,1], Patrice Péran [a,1]

[a] *ToNIC, Toulouse NeuroImaging Center, Université de Toulouse, Inserm, UPS, Toulouse, France*
[b] *LAAS CNRS, Université de Toulouse, CNRS, INSA, UPS, Toulouse, France*
[c] *CRCT, Centre de Recherche en Cancérologie de Toulouse, Inserm, UPS, Toulouse, France*

ARTICLE INFO

Keywords:
Brain MRI
Convolutional neural network
Deep learning
Diffusion-weighted imaging
Medical image classification
Simulated images

ABSTRACT

Convolutional neural networks (CNNs) are gradually being recognized in the neuroimaging community as a powerful tool for image analysis. Despite their outstanding performances, some aspects of CNN functioning are still not fully understood by human operators. We postulated that the interpretability of CNNs applied to neuroimaging data could be improved by investigating their behavior when they are fed data with known characteristics. We analyzed the ability of 3D CNNs to discriminate between original and altered whole-brain parametric maps derived from diffusion-weighted magnetic resonance imaging. The alteration consisted in linearly changing the voxel intensity of either one (monoregion) or two (biregion) anatomical regions in each brain volume, but without mimicking any neuropathology. Performing ten-fold cross-validation and using a hold-out set for testing, we assessed the CNNs' discrimination ability according to the intensity of the altered regions, comparing the latter's size and relative position. Monoregion CNNs showed that the larger the modified region, the smaller the intensity increase needed to achieve good performances. Biregion CNNs systematically outperformed monoregion CNNs, but could only detect one of the two target regions when tested on the corresponding monoregion images. Exploiting prior information on training data allowed for a better understanding of CNN behavior, especially when altered regions were combined. This can inform about the complexity of CNN pattern retrieval and elucidate misclassified examples, particularly relevant for pathological data. The proposed analytical approach may serve to gain insights into CNN behavior and guide the design of enhanced detection systems exploiting our prior knowledge.

1. Introduction

Convolutional neural networks (CNNs) have found great success in medical image analysis to perform a range of tasks, including classification and segmentation [1]. CNNs have shown promise when it comes to classifying neurological and neurodegenerative disorders [2,3] such as Alzheimer's disease (AD) [4] and Parkinson's disease (PD) [5]. These networks can directly process raw data freeing researchers from time-consuming manual feature extraction. Made up of multiple nonlinear modules that can create representations at simple yet abstract levels, CNNs automatically learn features during the training phase to optimize task resolution [6].

Advances in computational resources mean that 3D images can henceforth be used as input for CNNs. These have several advantages over 2D images [7–9]. In particular, 3D CNN architectures can integrate spatial information from the whole brain [10]. That represents a considerable advantage for data acquired with magnetic resonance imaging (MRI), which can provide structural and functional information about the entire brain volume [11].

Despite their remarkable performances, however, CNNs are regarded as black boxes, owing to their nontransparent decision-making process and difficult interpretation, sometimes hindering their usage [12,13]. To improve their interpretability, various techniques have been designed, such as producing explanations at the processing level (e.g. GradCAM [14], saliency maps [15]) or representations of different network components (i.e. layers, units), or creating self-explanatory models [16]. However, explanations remain marginal, as they always

[*] Corresponding author.
E-mail addresses: giulia-maria.mattia@inserm.fr (G.M. Mattia), villain.edouard@outlook.fr (E. Villain), federico.nemmi@gmail.com (F. Nemmi), mvlelann@laas.fr (M.V. Le Lann), xavier.franceries@inserm.fr (X. Franceries), patrice.peran@inserm.fr (P. Péran).
[1] These authors contributed equally.

https://doi.org/10.1016/j.artmed.2024.102897
Received 2 June 2023; Received in revised form 5 March 2024; Accepted 15 May 2024
Available online 17 May 2024

refer to CNN subparts [13]. As an alternative, we suggest shifting the focus to data, i.e. creating input data with specific characteristics to test CNN behavior, instead of solely considering architecture, learning rules, or objective functions [17].

Brain alterations in neurodegenerative diseases can be complex, involving several anatomical regions and pathophysiological changes [18]. Learning from neuropathological data to retrieve a comprehensive pathophysiological pattern may thus be extremely difficult as we cannot know to which extent individual patients' information contributes to this process. Inputting more homogeneous knowledge content into the network might shed light on how different features (e.g. the involved brain regions and their relative characteristics) influence CNN performance.

The present study aimed to ascertain whether we can study CNN behavior according to the provided input. More specifically, we modified brain MRI data to evaluate the discrimination ability of the proposed 3D CNN according to changes in the input, i.e. mean diffusivity (MD) parametric maps. To this end, we made specific alterations to the intensity of brain MRI data for two anatomical brain regions featured in mean diffusivity (MD) maps.

MD maps are computed from diffusion-weighted imaging (DWI), commonly used to extract parameters relating to the Brownian motion of water molecules [19]. Compared with other MRI indices, MD maps have the advantage of expressing a quantitative parameter (measured in mm^2/s) that corresponds to the mean voxelwise diffusion of water molecules [20,21].

Increase in MD values related to pathophysiological changes have already been observed in AD [22], PD [23] and multiple system atrophy (MSA) [24]. Therefore, we altered the original parametric maps (OPMaps) by linearly increasing the MD values of two specific brain regions: the cerebellum and putamen. These regions have highly dissimilar characteristics, in terms of tissue composition, shape, location and size. They are also affected in several neurodegenerative diseases, such as PD and MSA [24–28]. We were thus interested in investigating how even nonpathological modifications to these structures might impact CNN performance.

Although the alterations featured in the altered parametric maps (APMaps) were realistic compared to what can be observed in MD maps as a result of microstructural anomalies, they represented a plausible general pathological trait rather than reproducing a particular neuropathology.

To further explore the influence of brain region characteristics, we accounted for the different sizes of the two anatomical regions under consideration by extending or reducing the number of modified voxels while retaining their natural position. This size harmonization was done merely to analyze CNN behavior when confronted with regions of comparable size, with no intention of mimicking atrophy or other pathological conditions. This approach enabled us to establish whether the position of the target region inside the brain related to the number of modified voxels (i.e. region size) could impact CNN performance.

The alterations featured in the APMaps allowed us to establish a ground-truth behavior reflecting the discrimination ability of 3D CNNs when dealing with region-specific brain MRI parametric maps.

Moreover, we extended our contribution by creating biregion APMaps, namely APMaps in which both brain regions had been modified. Even though these modifications were well known from the user's point of view, we wished to show that CNN pattern retrieval is not straightforward when more than one altered region is present in the input data.

The aim of this study was to describe an approach that might facilitate the interpretation of 3D CNNs applied to brain MRI parametric maps, by inputting known data for the network to learn from. We postulated that if the input has known characteristics, results are more predictable and interpretable, given that we can anticipate what the CNN should look for.

To accomplish this aim, we established the regions of interest and relative features to test (intensity, position, and number of modified voxels), along with the architecture of the deep learning method (i.e. 3D CNN), and tracked the latter's performance. In recent work, we tested 3D CNNs trained to distinguish OPMaps from APMaps on an unseen set of 29 patients with MSA and 26 age-matched controls [29]. Performances were comparable to those of the state-of-the-art for differentiating patients with MSA from controls, proving the value of using APMaps to teach CNNs to recognize specific traits [29].

Basing the choice of the relevant parameters (e.g. the type of alteration, MRI modality, regions involved) on a priori knowledge concerning a specific pathology, we could investigate the discrimination abilities of different deep-learning approaches and select the most suitable one. This will ultimately improve the detection of the pathology of interest.

2. Materials and methods

This section describes our approach based on the creation of the APMaps (Section 2.1.3) and their successive exploitation as input data to a 3D CNN for studying its discrimination ability (Section 2.3). We aimed to evaluate CNN performance according to the different modifications introduced into the APMaps, i.e. changing the size and intensity of one or two brain regions.

2.1. Dataset

2.1.1. Participants and MRI protocol

A total of 89 participants (100% male) underwent brain imaging in a 3T MRI scanner (Philips Achieva) with a 32-channel head coil at the INSERM/UPS UMR1214 ToNIC technical platform (Toulouse, France). The mean age of the participants was 56.19 years (SD = 18.08, range = 20.67–85.25).

DWI acquisition parameters were as follows: TE = 55 ms; TR = 12.36 s; flip angle = 90°; FOV = 112 × 112 voxels; number of slices = 65; voxel size = 2 × 2 × 2 mm^3; EPI factor = 59; parallel factor = 2; phase encoding direction = postero-anterior; *b* value (number of directions) = 0 (1), 1000 (32) s/mm^2; total acquisition time = 16 min.

This study was approved by the local ethics committee and was conducted in accordance with the Declaration of Helsinki. Written informed consent was obtained from all participants. Additional information can be found in previous work [30].

2.1.2. Image preprocessing

Diffusion-weighted images were processed with the standard FSL pipeline [31], using fsl 5.0 [32]. This pipeline includes brain extraction, correction for eddy current and realignment, and fitting a standard tensor model to calculate MD (a more detailed description has been provided in a previous study [33]). We computed mean diffusivity maps and registered them in Montreal Neurological Institute (MNI) space with 3 × 3 × 3 mm^3 resolution by using nonlinear registration.

2.1.3. Creation of altered parametric maps

We developed a method for modifying MRI parametric maps of healthy brains by introducing region-specific alterations. To this end, we applied a linear intensity-based transformation to specific brain regions in the MD maps.

This straightforward variation in voxel intensity was in line with the physical meaning of MD maps, in that increased MD values generally indicate water diffusion anomalies, suggesting reduced microstructural integrity [19].

We selected the cerebellum and putamen as anatomical regions of interest, as they differ in four main respects:

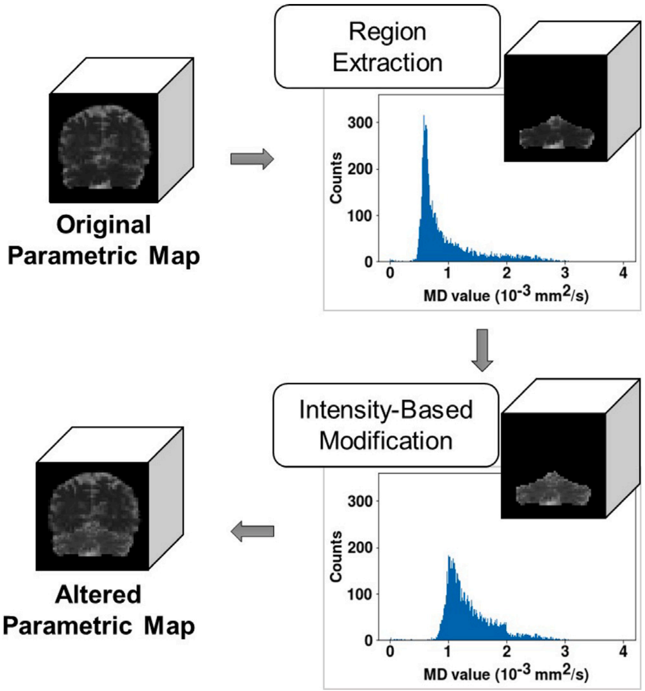


Fig. 1. Schematic representation of the method for creating altered brain MRI parametric maps (APMaps). The region of interest is extracted from the original parametric map, i.e. a healthy subject's mean diffusivity (MD) map, using an atlas. The altered parametric map is obtained by linearly increasing the MD values of the considered region, leaving the rest of the image unchanged.

- Position. The cerebellum is located underneath the brain hemispheres, surrounded by gray matter dorsally (the occipital lobe) and by the meninges and cerebrospinal fluid ventrally and posteriorly. By contrast, the putamen is located at the base of the forebrain, surrounded mainly by white matter;
- Morphology. The cerebellum is a single rounded structure, whereas the putamen, though rounded, is bilateral;
- Size. In normal individuals, the cerebellum has a mean volume of 300 cm^3 [34], whereas the putamen has a mean volume of around 3.60 cm^3 [35];
- Tissue composition. There is considerable heterogeneity in the cerebellum, owing to gray and white matter presence along with cerebrospinal fluid. The putamen is essentially composed of gray matter tissue.

In addition, the cerebellum and putamen are key regions for the assessment of a plethora of brain diseases, including motor disorders and cognitive dysfunctions [36–38]. For example, cerebellar ataxia and putaminal alterations are both encountered in MSA, categorized as an atypical Parkinsonian syndrome [24–28], and biomarkers involving the putamen are currently under development to distinguish Parkinson's disease from atypical syndromes [39].

The main steps for creating the APMaps are summarized in Fig. 1.

Regions of interest were extracted from the brain MRI volumes of normal individuals using an atlas [40]. Only voxels within these regions underwent the intensity modification, leaving the rest of each image unaltered.

The modifications to regional intensity were modeled as in Eq. (1), where $y_{r,n}$ and $x_{r,n}$ represent the altered and original regions (r), whose MD values lie below the nth intensity percentile (n), and p indicates the intensity increase as a percentage. Percentages ranged from 3% to 99%, in increments of 3%.

$$y_{r,n} = (1 + p) \cdot x_{r,n} \tag{1}$$

We chose either the 75th, 90th, or 100th percentile to limit image saturation effects. The 75th percentile was selected for the cerebellum, and the 90th for the putamen (for additional details, see Section S1.A, Supplementary Material).

Concerning size harmonization, we modified the size of each region by performing morphological operations on the respective atlas-based masks. Our goal was to determine whether the position of the region in relation to the number of modified voxels affected CNN performance.

We implemented the following morphological operations on the region masks:

- Erosion of the cerebellum (E-Cerebellum), to reach a size comparable to that of the putamen (about 400 voxels, given our resolution in MNI space);
- Dilation of the putamen (D-Putamen), to approximately match the size of the cerebellum (about 7200 voxels, given our resolution in MNI space).

These changes in region size served solely to establish a fair comparison to the anatomical reference and were not intended to imitate any pathological traits. The size harmonization process and examples of APMaps are given in Fig. 2.

We produced both monoregion APMaps, where only one region was modified in intensity, and biregion APMaps, where two regions were modified in intensity. Biregion APMaps are described in detail in Section 2.3.

2.2. Convolutional neural networks

CNNs allow for automatic feature extraction from multiple arrays (e.g. 3D images) and usually include a multilayer artificial neural network for classification tasks [6]. Interested readers can find additional information on CNNs in Section S1.B, Supplementary Material.

In the present study, we devised a 3D CNN for supervised binary classification, the task being to distinguish OPMaps from APMaps. Using the entire brain volume as CNN input preserves the spatial information of the whole MRI at a 3D participant level [41].

An overview of the proposed deep learning approach is provided in Fig. 3.

The CNN received as input the images (i.e. OPMaps and APMaps) in the shape of (60, 72, 60) voxels. Given the limited sample size, we carried out cross-validation as customary in the neuroimaging field [41, 42]. Each dataset (i.e. 89 paired images) was randomly divided: 80% for training with a ten-fold cross-validation and the remainder to serve as a hold-out set to assess CNN performance in the testing phase. The random seed for cross-validation was kept constant. Data normalization was performed on the training, validation, and hold-out sets by considering the maximum value of the training set for each fold. The best-epoch model with minimum loss value on the validation set was selected to establish network parameters and be tested on the hold-out set.

Our 3D CNN architecture was inspired by AlexNet [43] and VGG-Net [44]. Fig. 4 presents a schematic diagram of our model, comprising the following building blocks:

- *ConvBlock*, composed of a convolutional layer characterized by filter size = 3 × 3 × 3, stride = 1, with an increasing number of kernels going deeper into the network, and a batch normalization (BN) layer to speed up learning through a reduction in internal covariate shift [45], followed by an exponential linear unit (ELU) as the activation function [46];
- *Average Pooling*, to retain as much information as possible throughout the network, with filter size = 2 × 2 × 2 and stride = 2;
- *d-FC Block*, including a fully connected layer (FCL) with 512 neurons to ensure that enough units were available for the final classification, followed by a BN layer, an ELU activation, and a dropout layer, as part of a regularization technique intended to prevent overfitting [47];

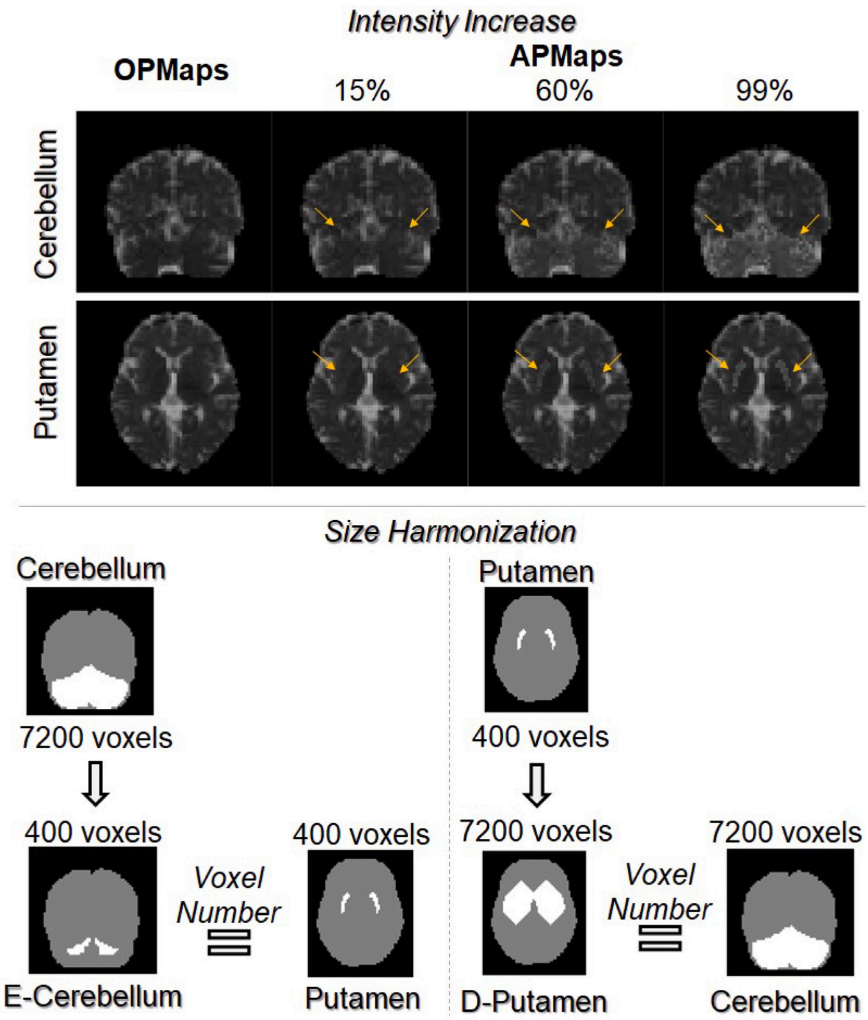


Fig. 2. Intensity increase and size harmonization. Top: Examples of APMaps according to increase in intensity as a percentage. Arrows point to the altered regions. Bottom: Size harmonization for the regions of interest with the corresponding number of voxels in each region mask. The brain is displayed in gray and the relevant region is in white. APMaps: Altered Parametric Maps; D: Dilated; E: Eroded; OPMaps: Original Parametric Maps.

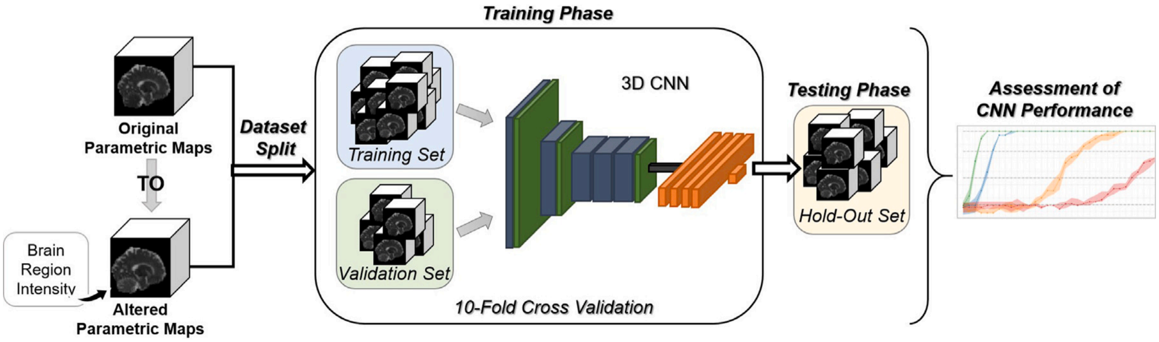


Fig. 3. Main steps of the proposed approach. MRI parametric maps of healthy individuals were modified to create the altered parametric maps by making intensity-based modifications to specific regions of interest. The original and altered parametric maps were split into three nonoverlapping sets: a training set and a validation set, derived from a ten-fold cross-validation scheme, and a hold-out set for the testing phase. A 3D convolutional neural network (CNN) was implemented to perform a binary classification task: original vs. altered parametric maps. The alterations made to the original parametric maps helped to assess how CNN performance varied according to changes in the input data.

- *FC Block*, same as *d-FC Block*, but without dropout;
- *FCL*, fully connected layer for binary classification with two neurons, followed by the softmax activation function.

The model was implemented using Keras library version 2.2.4 [48] and TensorFlow library version 1.13.1 [49] in Python version 3.6.9, supported by an NVIDIA® Quadro RTX™ 6000 graphical processing unit.

L2 regularization was applied with a factor of 0.0005 along with the *valid* method in convolutional layers to avoid padding [48]. The model trained over 100 epochs to prevent overfitting, with an initial learning rate of 0.00005, subject to dynamic reduction if there was no improvement in performance after five epochs. The training was

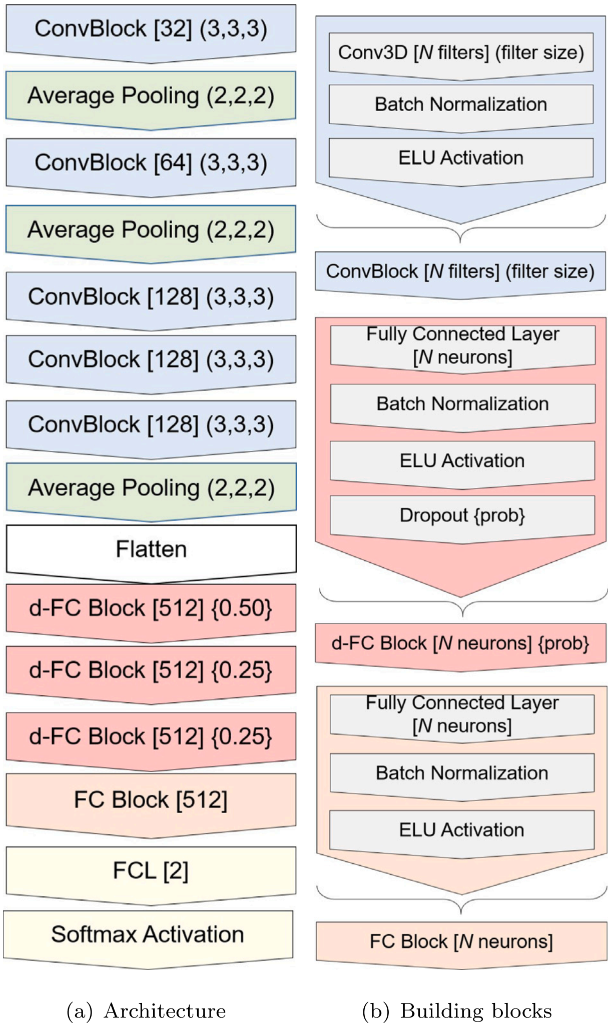


Fig. 4. Schematic diagram of the proposed 3D CNN (a) and its building blocks (b). The flatten operation yielded a 1-dimensional array for inputting to the FC layers. BN: Batch Normalization; CNN: Convolutional Neural Network; ELU: Exponential Linear Unit; FC: Fully Connected; FCL: Fully Connected Layer; prob: dropout probability.

carried out using mini-batch gradient descent, with a batch size of eight samples to meet computational requirements. Categorical cross-entropy (i.e. logarithmic loss function) was used with Adam optimizer. This optimizer is characterized by the combination of an adaptive gradient algorithm with root mean square propagation and is especially suited to problems involving large amounts of data [50].

The code for creating the APMaps and the CNN model has been released in a GitHub repository [51].

For model evaluation, we computed accuracy, measuring the overall performance of the network, sensitivity and specificity, indicating respectively the proportion of APMaps and OPMaps correctly classified as such. We provide exhaustive definitions in Section S1.C, Supplementary Material. All metrics are presented as the median and the interquartile range (IQR) obtained on the hold-out set, given the adoption of a ten-fold cross-validation.

2.3. Experiments

CNN performance was first assessed using OPMaps and monoregion APMaps as input, with intensity increases between 3% and 99%, in increments of 3%, for each region.

Based on these results, we established four levels of accuracy: very low (VL), low (L), fair (F), and high (H), with reference values of 0.45, 0.65, 0.85, and 1.00. To create the biregion APMaps, we combined regions according to their size and the accuracy levels achieved by the CNN trained with the respective monoregion APMaps. For brevity's sake, we defined monoregion-trained and biregion-trained CNNs according to the input data used in the performance assessment (i.e. monoregion or biregion APMaps together with OPMaps). Biregion APMaps featured two modified regions, which were paired according to their size: either different (i.e. Cerebellum/Putamen) or comparable (i.e. D-Putamen/Cerebellum and E-Cerebellum/Putamen). Creating biregion APMaps allowed us to increase the complexity of the input data, thereby approaching realistic pathological conditions where more than one region is altered while keeping the training content still known.

Monoregion-trained CNNs were associated with one of the four accuracy levels depending on the achieved accuracy values. If needed, additional intensity increases were computed in 1% increments to match the accuracy levels as closely as possible. When the same accuracy value (e.g. equal to 1.00) corresponded to different intensity increases, we selected the one with the highest minimum accuracy across the ten folds presenting the lowest intensity increase.

Biregion APMaps were obtained by applying the method described in Section 2.1.3 with the intensity increase corresponding to the accuracy level and region dictated by the monoregion-trained CNNs (see Table S1, Supplementary Material).

To evaluate the relative effects of position and size, we examined all 16 possible combinations of accuracy levels, either the same (i.e. VL/VL, L/L, F/F, H/H) or different (e.g. VL/L, L/F) between regions.

Moreover, we compared monoregion-trained CNNs with biregion-trained CNNs by testing monoregion-trained CNNs on biregion APMaps and vice versa to find out the contribution of each accuracy level to the learned patterns. To this end, we examined the following cases:

- CNNs trained with biregion APMaps with the H/H accuracy combination and tested on monoregion APMaps with intensity increases dictated by the corresponding H accuracy level;
- CNNs trained with monoregion APMaps with intensity increases dictated by the H accuracy level and tested on the corresponding biregion APMaps with the H/H accuracy combination.

3. Results

3.1. Monoregion-trained CNNs

The investigation of CNN behavior began by assessing the ability to distinguish between OPMaps and monoregion APMaps modified across a range of intensity increases. Fig. 5 shows the median accuracy and IQR achieved on the hold-out set for each intensity increase applied to the regions.

Cerebellum and D-Putamen CNNs exhibited similar behavior, even though the former reached maximum accuracy with a higher intensity increase than the latter (27% vs. 15%).

The Putamen CNN achieved an accuracy of 1.00 at 84%, whereas the E-Cerebellum CNN only reached an accuracy of 0.81 with an intensity increase of 99%. Despite comparable region size, the E-Cerebellum CNN only overcame near-to-chance accuracy with a 75% intensity increase (vs. 45% for the Putamen CNN).

Regarding sensitivity and specificity, we found overall comparable behavior with respect to accuracy, i.e. increasing performances with the percentage used to modify the intensity of each brain region in the APMaps. However, a higher IQR can be observed for the sensitivity over low-intensity increases, although less evident for the D-Putamen and Cerebellum CNNs (intensity increase <12%) and the Putamen CNN (intensity increase <50%). The sensitivity of the E-Cerebellum

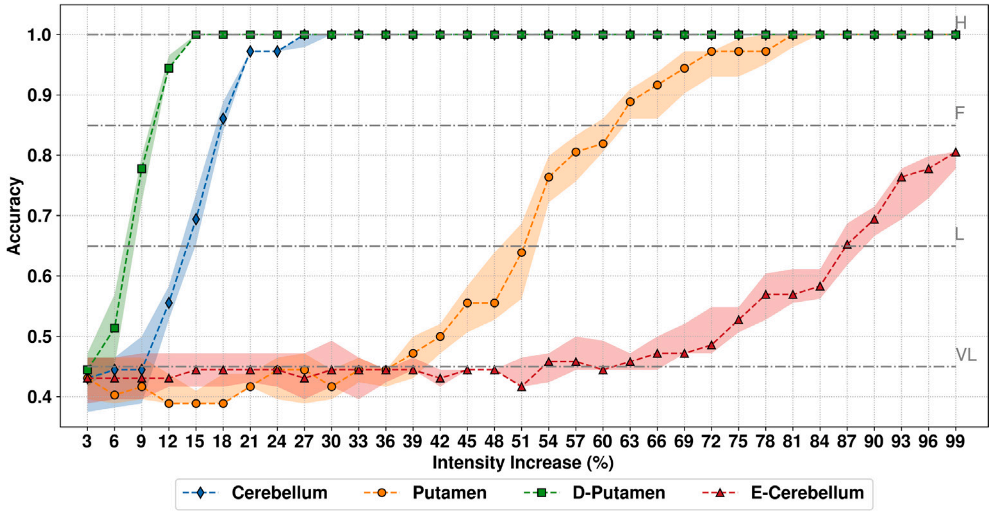


Fig. 5. Performance of the monoregion-trained CNNs. Accuracy on hold-out set given as median and IQR over ten-fold cross-validation according to the intensity increase in the APMaps. Gray lines indicate the four accuracy levels used for performance assessment. APMaps: Altered Parametric Maps; CNN: Convolutional Neural Network; D: Dilated; E: Eroded; F: Fair; H: High; IQR: Interquartile Range; L: Low; VL: Very Low.

CNN revealed more fluctuations in the median value and higher IQR until reaching intensity increases over 90%. The specificity presented fewer variations and lower IQR compared to the sensitivity across all monoregion-trained CNNs. These results are available in Section S2.A, Supplementary Material.

3.2. Biregion-trained CNNs

Biregion-trained CNNs had to distinguish OPMaps from biregion APMaps featuring two regions modified in intensity. Bar plots in Fig. 6 represent the accuracy achieved by the biregion-trained CNNs compared with the best accuracy reached by the monoregion-trained CNNs for the different combinations of accuracy levels. Unpaired Student *t* tests computed between biregion accuracy and the best monoregion accuracy showed that biregion performance was significantly better for most of the VL/VL, L/L, and F/F combinations. For mixed combinations of F, L, and VL levels, biregion-trained CNNs significantly outperformed their monoregion counterparts (e.g. for VL/L, L/F). No significant difference was found only for the combinations VL/VL and F/VL of the D-Putamen/Cerebellum CNN and VL/F of the Cerebellum/Putamen CNN. All three pairs of regions showed excellent performance (accuracy equal to 1.00) when at least one of the two regions was characterized by the H accuracy level.

Using one-way analysis of variance (ANOVA), we identified meaningful differences in performances between the combinations of levels of accuracy for the pairs represented by blue, orange, or green bars in Fig. 6. Accuracy was significantly lower for VL/VL than for the other combinations, with all the other comparisons (e.g. VL/L vs. L/VL, VL/F vs. F/VL) revealing smaller differences in accuracy.

For clarity's sake, significant differences derived from the one-way ANOVA are not specified in Fig. 6, but are listed in Table S2, Supplementary Material.

Findings regarding sensitivity and specificity obtained for each accuracy level did not differ much from what emerged for accuracy, remaining coherent for most comparisons (see Section S2.B, Supplementary Material).

3.3. Monoregion- vs. Biregion-trained CNNs

We tested monoregion-trained CNNs on their ability to distinguish OPMaps from biregion APMaps. Similarly, biregion-trained CNNs had to distinguish OPMaps from monoregion APMaps, insofar as the intensity increases between regions in the monoregion and biregion APMaps were the same.

We can see from Table 1 that monoregion-trained CNNs successfully classified the altered target regions in each biregion image, achieving the highest performance for each of them.

By contrast, biregion-trained CNNs performed poorly on some of the respective monoregion testing images: D-Putamen/Cerebellum CNN trained with the D-Putamen/Cerebellum APMaps successfully classified the Cerebellum APMaps achieving a median accuracy of 0.89 but performed poorly on the D-Putamen APMaps. The biregion CNN trained with E-Cerebellum/Putamen APMaps achieved a median accuracy of around 0.65 for both the Putamen and E-Cerebellum APMaps.

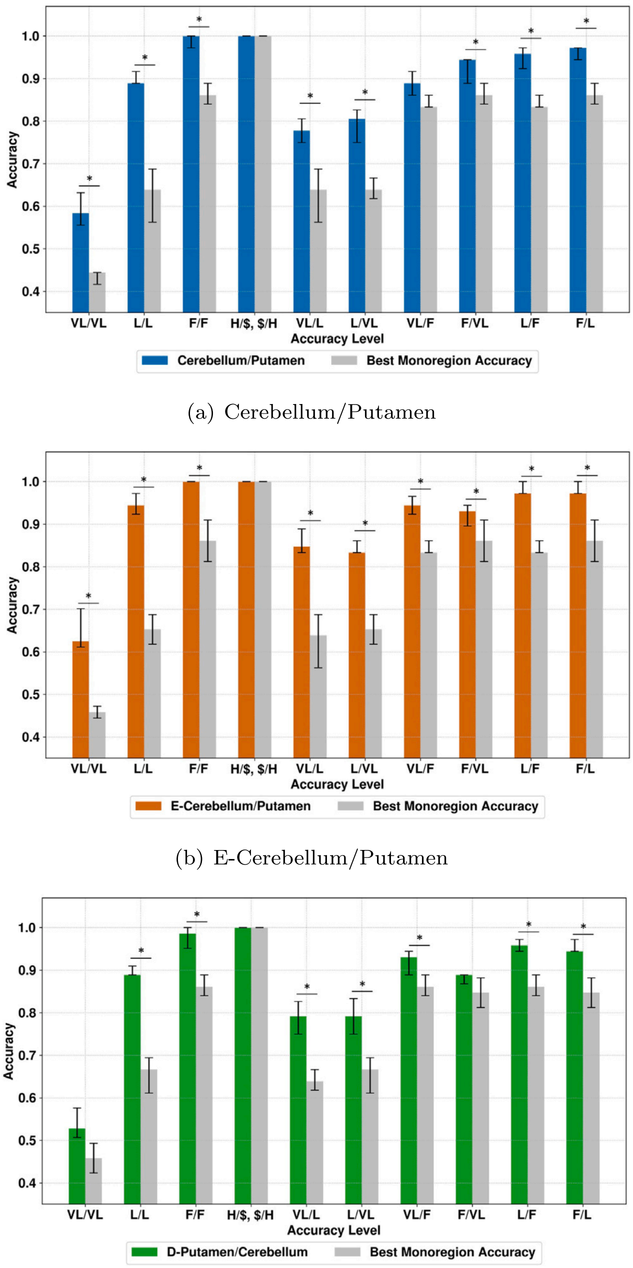
Regarding regions of different sizes, the Cerebellum/Putamen CNN recognized Cerebellum APMaps almost perfectly, with a median accuracy of 0.97, despite the much inferior performance obtained on the Putamen APMaps, with a median accuracy of 0.50. Overall, the OPMaps were correctly classified by the biregion-trained CNNs (specificity equal to 1), whereas the ability to recognize properly the monoregion APMaps varied according to the involved region. The Cerebellum/Putamen CNN reported the highest sensitivity (0.94) on the Cerebellum APMaps and the lowest on the Putamen APMaps.

4. Discussion

In the present study, we analyzed the behavior of a 3D CNN when fed with ad hoc modified brain MRI parametric maps.

As we had planned the modifications made to the data, we knew what content was provided to the CNN, guiding us to interpret and track the changes in performance according to the input characteristics. This process would not have been so straightforward had we used a typical set of pathological data, as these inevitably include unknown components, owing to the different ways each disease can manifest itself. We can assume that pattern retrieval becomes more difficult as the heterogeneity of the data associated with each label increases.

To the best of our knowledge, this is the first study to have fed a CNN with 3D neuroimaging data altered in a realistic and controlled manner, i.e. modifying the value of specific brain regions according to the physical meaning of the relevant MRI index. The results we obtained were easier to interpret and helped us understand how CNN behavior changes according to specific input features, such as intensity, which is sensitive to modifications in a variety of pathologies. In line with our expectations, we found that the larger the region, the smaller



(a) Cerebellum/Putamen

(b) E-Cerebellum/Putamen

(c) D-Putamen/Cerebellum

Fig. 6. Performance of the biregion-trained CNNs. Median accuracy and IQR on hold-out set over a ten-fold cross-validation compared with the best performance of monoregion-trained CNNs considering 16 combinations of accuracy levels. The dollar sign stands for VL, L, F, H, as all combinations featuring at least one H resulted in equal performances. $*$ $p < 0.05$. CNN: Convolutional Neural Network; D: Dilated; E: Eroded; F: Fair; H: High; IQR: Interquartile Range; L: Low; VL: Very Low. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

the intensity increase required to achieve good performances by the monoregion-trained CNNs.

In addition, we investigated the influence of the position of the altered regions inside the brain by creating the D-Putamen and E-Cerebellum APMaps. These contained a comparable number of modified voxels to the cerebellum or the putamen.

We found that as more centrally located, the putamen and D-Putamen outperformed their equally sized but more peripheral counterparts (i.e. the E-Cerebellum and cerebellum). Therefore, it seemed easier for our 3D CNN to detect regions in the center of the brain rather than on its periphery (e.g. comparing the performance of D-Putamen CNN vs. Cerebellum CNN in Fig. 5). Nonetheless, additional experiments may help clarify this point.

Moreover, we could determine an intensity threshold for each brain region to ensure that the 3D CNN achieved high accuracy, as suggested by the accuracy plateaus equal to 1.00 in Fig. 5. Considering the sensitivity, monoregion CNNs showed greater instability on the performance of smaller regions, i.e. E-Cerebellum and putamen, especially for low-intensity increase. Smaller regions with less intense modifications seemed to complicate the classification task for the CNN. Instead, the CNN ability to classify OPMaps grew stably with the intensity increase, as did the accuracy, albeit with higher variability regarding smaller regions. The more heterogeneous nature of the APMaps due to the alteration in intensity absent from the OPMaps may explain this behavior. These findings concretely demonstrate CNN sensitivity to both the intensity and position of the modified regions and the extent to which these impacted pattern retrieval.

In a recent study, we used 3D CNNs trained with APMaps and OPMaps to differentiate between 29 patients with MSA and 26 age-matched controls [29]. Performances were in accord with the state-of-the-art for MSA classification, proving that the traits learned from the APMaps enclosed salient features that could also be detected in the unseen brain MD maps of patients with MSA. This approach offers a way of coping with small sample sizes in the case of rare diseases such as MSA, making it feasible to use deep learning by exploiting a priori knowledge of the disease. In another recent work [52], a deep learning fused model based on diffusion-weighted images and clinical data allowed for predicting the functional outcome of acute ischemic stroke patients. The automatic feature extraction from DWI and B0 images via a 3D CNN enabled lesion characterization in a data-driven manner, e.g. incorporating position and tissue variability, which can be difficult to express with clinical variables. When dealing with mild stroke severity, the performance was slightly inferior compared to more severe conditions. These results may align with ours in that detecting milder modifications can be more challenging.

Using the results obtained with the monoregion-trained CNNs as our baseline, we examined how the CNN behaved when fed with APMaps featuring two altered brain regions. We created biregion APMaps based on four levels of accuracy (VL, L, F, and H, from lowest to highest) exhibited by the monoregion-trained CNNs and combined the regions accordingly.

The knowledge content corresponding to the combination of two accuracy levels significantly improved the performances of the biregion-trained CNNs, compared with when the regions were considered on their own (see L/L, VL/L, etc., in Fig. 6). This result is encouraging insofar as each region considered separately provided insufficient information, whereas when the regions were combined, there was enough information to provide a detectable pattern. Most remarkable was the case of the L/L biregion-trained CNNs, which exceeded the accuracy of the monoregion-trained CNNs by at least 20%, with a final accuracy around 0.90. Unsurprisingly, when at least one region was featured with the H accuracy level, biregion-trained CNNs yielded equally excellent performances independently of region size. No difference was found between the mixed accuracy combinations (e.g. VL/L vs. L/VL), suggesting that the content provided by each accuracy level was enough to boost performance regardless of the region's characteristics.

The last part of our investigation was designed to ascertain whether a biregion-trained CNN could detect abnormal traits in a single region. If pattern retrieval were simply additive, in terms of the information provided to the CNN, we would expect the latter to be capable of detecting each altered region individually.

In the case of monoregion-trained CNNs tested on the biregion APMaps, we found that they performed well on each region, as can be observed from Table 1. By contrast, when we used monoregion APMaps to test biregion-trained CNNs, the latter performed more poorly than

Table 1

Monoregion- vs. Biregion-Trained CNNs: CNNs trained with monoregion APMaps matching the H accuracy level were tested on the corresponding H/H biregion hold-out set and vice versa. Metrics are provided as median (IQR) over a ten-fold cross-validation. APMaps: Altered Parametric Maps; CNN: Convolutional Neural Network; D: Dilated; E: Eroded; H: High; IQR: Interquartile Range.

Training	*Testing*			
Monoregion-trained CNN	**Biregion APMaps**	**Accuracy**	**Sensitivity**	**Specificity**
Cerebellum	Cerebellum/Putamen	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)
Putamen		1.00 (0.00)	1.00 (0.00)	1.00 (0.00)
E-Cerebellum	E-Cerebellum/Putamen	1.00 (0.03)	1.00 (0.06)	1.00 (0.00)
Putamen		1.00 (0.00)	1.00 (0.00)	1.00 (0.00)
D-Putamen	D-Putamen/Cerebellum	1.00 (0.03)	1.00 (0.06)	1.00 (0.00)
Cerebellum		1.00 (0.00)	1.00 (0.00)	1.00 (0.00)
Biregion-trained CNN	**Monoregion APMaps**	**Accuracy**	**Sensitivity**	**Specificity**
Cerebellum/Putamen	Cerebellum	0.97 (0.02)	0.94 (0.04)	1.00 (0.00)
	Putamen	0.50 (0.02)	0.00 (0.04)	1.00 (0.00)
E-Cerebellum/Putamen	E-Cerebellum	0.64 (0.08)	0.28 (0.17)	1.00 (0.00)
	Putamen	0.65 (0.03)	0.31 (0.06)	1.00 (0.00)
D-Putamen/Cerebellum	D-Putamen	0.56 (0.08)	0.11 (0.15)	1.00 (0.00)
	Cerebellum	0.89 (0.10)	0.78 (0.21)	1.00 (0.00)

the relevant monoregion-trained CNN on at least one of the two regions. Monoregion- and biregion-trained CNNs presented all maximum specificity but varied in sensitivity, proving that the positive class, i.e. the APMaps, carried somewhat different information compared to the training content. One possible explanation is that the biregion-trained CNNs learned a multispatial signature, which was absent from the monoregion APMaps, hence the unsatisfactory performances on most regions considered singly. Given the results in Table 1, we can say that either the larger yet less intense region (i.e. cerebellum in the Cerebellum/Putamen CNN), or the more intense yet more peripheral region (i.e. cerebellum in the D-Putamen/Cerebellum CNN) was well detected (more details in Table S1, Supplementary Material). This suggests that in the presence of more than one altered region, more importance may be given to specific characteristics of the altered regions (e.g. intensity, size, and position) during the feature extraction process and/or the final classification.

These findings could inform clinical research. Neurodegenerative lesions start in one specific site and spread as the disease progresses [18]. Consequently, monoregion APMaps may represent early pathological states and biregion APMaps more advanced states. CNNs trained with data derived from early pathological conditions (e.g. monoregion APMaps) may also be able to classify advanced stages (e.g. biregion APMaps) involving the same region. Using biregion APMaps to classify earlier pathological conditions may be less effective, depending on the region of interest, as the learned patterns may carry a multispatial signature incompatible with an earlier stage of the disease. Therefore, building region-specific CNNs may ameliorate detection accuracy for abnormalities in a particular brain region. The successive merging of these results, for instance, by using majority voting or other techniques, may eventually boost the overall performance. In this regard, the modifications applied to the brain regions studied here were oversimplified compared to the complexity of neuropathological patterns. We aimed to ascertain how the interpretation of CNN behavior is affected when we operators are well-informed about the data these systems learn from and how we can use this advantage fully to our benefit.

One limitation of the present study was the small sample size, related to the use of real-world neuroimaging data. Nevertheless, modifying brain parametric maps of healthy individuals across a wide age range enabled us to preserve the interindividual variability and intrinsic heterogeneity in terms of morphology and anatomy. The advantage here lies in reducing the distortions or artifacts that may arise from a completely artificial set [53,54]. Although moderate, this sample size allowed for testing on an external unseen set of pathological data as demonstrated in a previous study [29], reaching a good generalization performance (accuracy higher than 0.8). In future work, we plan to extend the sample size, including female subjects as well, and evaluate the impact of increasing sample size with specific input characteristics on the network's performance.

Another deep learning approach, namely autoencoders for unsupervised anomaly detection (UAD), relying on the images of healthy controls for training, has been burgeoning with promising results, e.g. for the detection of brain tumors or multiple sclerosis lesions [55,56]. This unsupervised strategy can cope with data scarcity encountered in the medical domain, as no annotation is needed for training [57]. Briefly, anomalies in a patient population can result from the poor reconstruction provided by the autoencoder, as it learned to model the distribution of the healthy brain from the images. A recent study explored autoencoders to detect subtle anomalies from brain diffusion MRI of 129 de novo PD patients proposing different autoencoders trained on a set of 56 healthy controls [58]. Overall, diffuse cerebral anomalies were revealed rather than finding a specific biomarker for early PD. However, the reconstruction ability of the model seemed poorer for small regions, such as the substantia nigra, which is of great interest in PD pathophysiology. A way to combine the strength of the latter and our approach would be to create altered parametric maps targeting the substantia nigra and explore the reconstruction ability of the autoencoder to determine a sensitivity threshold. Compared with UAD approaches, our goal was to better interpret supervised CNN behavior thanks to prior knowledge of the data, whereas UAD does not require a ground truth to work well, being unsupervised, but could reveal new information about the data. Although different, both approaches could benefit from these complementary aspects to reach maximum potential. In this regard, a recent study on anomaly detection and segmentation tasks exploited real and synthetic data to test the capacity of variational autoencoders and transformers applied to 2D and 3D medical images [59]. Remarkably, performances dropped when the intensity modification of the synthetic images got closer to tissue characteristics. These experiments can thus facilitate our interpretation of deep learning methods applied to supervised and unsupervised settings.

It goes without saying that the results discussed in this study are closely related to the 3D CNN architecture we adopted and the brain regions we chose when modifying the MRI data. However, this encourages us to explore a variety of settings, now that we have established a baseline. Indeed, beginning from prior knowledge about a particular pathological condition, we could explore any CNN architecture's discrimination capacity by feeding more targeted APMaps.

5. Conclusion

In the present study, we investigated the discrimination ability of a 3D CNN to distinguish original from altered whole-brain MRI parametric maps. By linearly modifying the intensity of one (monoregion) or two (biregion) brain regions, i.e. the cerebellum and putamen, we showed how salient input features, such as size, position, and intensity, influence CNN performance. Although these alterations were independent of any specific neuropathology, they were in line with the physical significance of the considered MRI index.

Monoregion-trained CNNs proved that the greater and more intense the modified region, the easier its discrimination. Results from biregion-trained CNNs were significantly better than those from their monoregion counterpart, pointing out the importance of the joined contribution of the altered regions rather than considering them alone.

Creating APMaps with different target regions and types of MRI data may help us to customize CNNs, and respond to specific concerns about why certain patterns are better discriminated than others, thanks to the ground truth constituted by the APMaps. These have already been used as ground truth images to assess the performance of a straightforward visualization technique for CNN interpretability in the case of 3D neuroimaging data, opening up the way for other possible uses [60].

Building on the present study, we intend to increase the input complexity by creating APMaps that evoke specific pathologies and evaluating how CNNs react to increasingly varied knowledge content. We hope our approach will pave the way for further applications comprising different deep learning architectures and regions of interest, possibly even beyond the brain, to favor the interpretability (and hence the use) of deep learning applied to biomedical data. These findings are just the starting point when it comes to grasping how the complexity of input data influences CNN pattern retrieval.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

CRediT authorship contribution statement

Giulia Maria Mattia: Writing – review & editing, Writing – original draft, Validation, Software, Methodology, Formal analysis, Data curation, Conceptualization. **Edouard Villain:** Writing – review & editing, Writing – original draft. **Federico Nemmi:** Writing – review & editing, Writing – original draft. **Marie-Véronique Le Lann:** Writing – review & editing, Writing – original draft. **Xavier Franceries:** Writing – review & editing, Writing – original draft, Supervision, Methodology, Conceptualization. **Patrice Péran:** Writing – review & editing, Writing – original draft, Supervision, Methodology, Conceptualization.

Declaration of competing interest

The authors declare no conflict of interest.

Data availability

The data that support the findings of this study may be available upon reasonable request.

Acknowledgments

We would like to thank the MRI technical platform at Toulouse Neuroimaging Center (ToNIC), Université de Toulouse, Inserm, UPS, France.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.artmed.2024.102897.

References

[1] Litjens G, et al. A survey on deep learning in medical image analysis. Med Image Anal 2017;42:60–88. http://dx.doi.org/10.1016/j.media.2017.07.005.

[2] Vieira S, Pinaya WHL, Mechelli A. Using deep learning to investigate the neuroimaging correlates of psychiatric and neurological disorders: Methods and applications. Neurosci Biobehav Rev 2017;74:58–75. http://dx.doi.org/10.1016/j.neubiorev.2017.01.002.

[3] Noor MBT, Zenia NZ, Kaiser MS, Mahmud M, Mamun SA. Detecting neurodegenerative disease from MRI: A brief review on a deep learning perspective. In: Lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics). LNAI, LNAI, 2019;Vol. 11976:115–25. http://dx.doi.org/10.1007/978-3-030-37078-7_12.URL https://link.springer.com/chapter/10.1007/978-3-030-37078-7_12,

[4] Payan A, Montana G. Predicting Alzheimer's disease: a neuroimaging study with 3D convolutional neural networks. 2015, arXiv:1502.02506, Preprint.

[5] Martínez-Murcia FJ, et al. A 3D convolutional neural network approach for the diagnosis of Parkinson's disease. In: Ferrández Vicente JM, Ramón A-SJ, de la Paz López F, Toledo Moreo J, Adeli H, editors. Natural and artificial computation for biomedicine and neuroscience. Cham: Springer International Publishing; 2017, p. 324–33. http://dx.doi.org/10.1007/978-3-319-59740-9_32.

[6] LeCun Y, Bengio Y, Hinton G. Deep learning. Nature 2015;521:436–44. http://dx.doi.org/10.1038/nature14539.

[7] Esmaeilzadeh S, Yang Y, Adeli E. End-to-end Parkinson disease diagnosis using brain MR-images by 3D-CNN. 2018, arXiv:1806.05233, Preprint.

[8] Sarraf S, DeSouza DD, Anderson J, Tofighi G, for the Alzheimer's Disease Neuroimaging Initiativ. DeepAD: Alzheimer's disease classification via deep convolutional neural networks using MRI and fMRI. bioRxiv 2017. http://dx.doi.org/10.1101/070441, URL https://www.biorxiv.org/content/early/2017/01/14/070441, Preprint.

[9] Hosseini-Asl E, Ghazal M, Mahmoud AH, Aslantas A, Shalaby AM, Casanova MF, Barnes GN, Gimel'farb GL, Keynton RS, El-Baz AS. Alzheimer's disease diagnostics by a 3D deeply supervised adaptable convolutional network. Front Biosci-Landmark (FBL) 2018;23(3):584–96. http://dx.doi.org/10.2741/4606.

[10] Trivizakis E, et al. Extending 2-D convolutional neural networks to 3-D for advancing deep learning cancer classification with application to MRI liver tumor differentiation. IEEE J Biomed Health Inf 2019;23:923–30. http://dx.doi.org/10.1109/JBHI.2018.2886276.

[11] Rosenbloom M, Pfefferbaum A. Magnetic resonance imaging of the living brain: Evidence for brain degeneration among alcoholics and recovery with abstinence. Alcohol Res Health 2008;31:362–76.

[12] Shrikumar A, Greenside P, Kundaje A. Learning important features through propagating activation differences. In: Proceedings of the 34th international conference on machine learning - volume 70. ICML '17, JMLR.org; 2017, p. 3145–53.

[13] Elton DC. Self-explaining AI as an alternative to interpretable AI. In: Goertzel B, Panov AI, Potapov A, Yampolskiy R, editors. Artificial general intelligence. Cham: Springer International Publishing; 2020, p. 95–106. http://dx.doi.org/10.1007/978-3-030-52152-3_10.

[14] Selvaraju RR, Das A, Vedantam R, Cogswell M, Parikh D, Batra D. Grad-CAM: Visual explanations from deep networks via gradient-based localization. Int J Comput Vis 2019;128:336–59. http://dx.doi.org/10.1109/ICCV.2017.74.

[15] Simonyan K, Vedaldi A, Zisserman A. Deep inside convolutional networks: Visualising image classification models and saliency maps. In: Bengio Y, LeCun Y, editors. 2nd international conference on learning representations, ICLR 2014, banff, AB, Canada, April 14-16, 2014, workshop track proceedings. 2014, URL http://arxiv.org/abs/1312.6034.

[16] Gilpin LH, Bau D, Yuan BZ, Bajwa A, Specter M, Kagal L. Explaining explanations: An overview of interpretability of machine learning. In: 2018 IEEE 5th international conference on data science and advanced analytics. DSAA, 2018, p. 80–9. http://dx.doi.org/10.1109/DSAA.2018.00018.

[17] Richards BA, et al. A deep learning framework for neuroscience. Nat Rev Neurosci 2019;22:1761–70. http://dx.doi.org/10.1038/s41593-019-0520-2.

[18] Brettschneider J, Tredici K, Lee V, Trojanowski J. Spreading of pathology in neurodegenerative diseases: A focus on human studies. Nat Rev Neurosci 2015;16:109–20. http://dx.doi.org/10.1038/nrn3887.

[19] Bihan D. Looking into the functional architecture of the brain with diffusion MRI. Nat Rev Neurosci 2003;4:469–80. http://dx.doi.org/10.1038/nrn1119.

[20] Kim H, Kim S, Kim HS, Choi C, Lee C. Alterations of mean diffusivity in brain white matter and deep gray matter in Parkinson's disease. Neurosci Lett 2013;550:64–8. http://dx.doi.org/10.1016/j.neulet.2013.06.050.

[21] Vos S, Jones D, Jeurissen B, Viergever M, Leemans A. The influence of complex white matter architecture on the mean diffusivity in diffusion tensor MRI of the human brain. NeuroImage 2012;59:2208–16. http://dx.doi.org/10.1016/j.neuroimage.2011.09.086.

[22] Eustache P, Nemmi F, Saint-Aubert L, Pariente J, Péran P. Multimodal magnetic resonance imaging in Alzheimer's disease patients at prodromal stage. J Alzheimer's Dis 2016;50:1035–50. http://dx.doi.org/10.3233/JAD-150353.

[23] Péran P, et al. Magnetic resonance imaging markers of Parkinson's disease nigrostriatal signature. Brain 2010;133(11):3423–33. http://dx.doi.org/10.1093/brain/awq212.

[24] Péran P, et al. MRI supervised and unsupervised classification of Parkinson's disease and multiple system atrophy. Mov Disord 2018;33:600–8. http://dx.doi.org/10.1002/mds.27307.

[25] Berg D, Steinberger J, Olanow CW, Naidich T, Yousry T. Milestones in magnetic resonance imaging and transcranial sonography of movement disorders. Mov Disord 2011;26(6):979–92. http://dx.doi.org/10.1002/mds.23766.

[26] Shin H, Kang S, Yang JH, Kim H, Lee M-S, Sohn Y. Use of the putamen/caudate volume ratio for early differentiation between parkinsonian variant of multiple system atrophy and Parkinson Disease. J Clin Neurol (Seoul, Korea) 2007;3(2):79–81. http://dx.doi.org/10.3988/jcn.2007.3.2.79.

[27] Seppi K, et al. Progression of putaminal degeneration in multiple system atrophy: A serial diffusion MR study. NeuroImage 2006;31:240–5. http://dx.doi.org/10.1016/j.neuroimage.2005.12.006.

[28] Barbagallo G, et al. Multimodal MRI assessment of nigro-striatal pathway in multiple system atrophy and Parkinson disease. Mov Disord 2016;31(3):325–34. http://dx.doi.org/10.1002/mds.26471.

[29] Mattia GM, Villain E, Nemmi F, Rascol O, Meissner WG, Franceries X, Péran P. Neurodegenerative traits detected via 3D CNNs trained with simulated brain MRI: Prediction supported by visualization of discriminant voxels. In: 2021 IEEE international conference on bioinformatics and biomedicine. BIBM, 2021, p. 1437–42. http://dx.doi.org/10.1109/BIBM52615.2021.9669894.

[30] Nemmi F, Levardon M, Péran P. Brain-age estimation accuracy is significantly increased using multishell free-water reconstruction. Hum Brain Mapp 2022;43:2365–76. http://dx.doi.org/10.1002/hbm.25792.

[31] Behrens T, et al. Characterization and propagation of uncertainty in diffusion-weighted MR imaging. Magn Reson Med 2003;50(5):1077–88. http://dx.doi.org/10.1002/mrm.10609.

[32] Jenkinson M, Beckmann CF, Behrens TE, Woolrich MW, Smith SM. FSL. NeuroImage 2012;62(2):782–90. http://dx.doi.org/10.1016/j.neuroimage.2011.09.015, URL https://www.sciencedirect.com/science/article/pii/S1053811911010603, 20 YEARS OF fMRI.

[33] Nemmi F, Pavy-Le Traon A, Phillips O, Galitzky M, Meissner W, Rascol O, Péran P. A totally data-driven whole-brain multimodal pipeline for the discrimination of Parkinson's disease, multiple system atrophy and healthy control. NeuroImage: Clin 2019;23:101858. http://dx.doi.org/10.1016/j.nicl.2019.101858, URL https://www.sciencedirect.com/science/article/pii/S2213158219302086.

[34] Shepherd G. The synaptic organization of the brain. New York: Oxford University Press; 2004, http://dx.doi.org/10.1093/acprof:oso/9780195159561.001.1.

[35] Yin D, Valles F, Fiandaca M, Forsayeth J, Bankiewicz K. Striatal volume differences between non-human and human primates. J Neurosci Methods 2009;176:200–5. http://dx.doi.org/10.1016/j.jneumeth.2008.08.027.

[36] Molinari M, Leggio M. Cerebellum: Clinical pathology. In: Encyclopedia of neuroscience. Elsevier Ltd; 2010, p. 737–42. http://dx.doi.org/10.1016/B978-008045046-9.00567-2.

[37] Viñas-Guasch N, Wu YJ. The role of the putamen in language: a meta-analytic connectivity modeling study. Brain Struct Funct 2017;222:3991–4004. http://dx.doi.org/10.1007/s00429-017-1450-y.

[38] Haber S. Corticostriatal circuitry. Dialogues Clin Neurosci 2016;18:7–21. http://dx.doi.org/10.31887/DCNS.2016.18.1/shaber.

[39] Michell A, Lewis S, Foltynie T, Barker R. Biomarkers and Parkinson's disease. Brain 2004;127 Pt 8:1693–705. http://dx.doi.org/10.1093/brain/awh198.

[40] Hammers A, et al. Three-dimensional maximum probability atlas of the human brain, with particular reference to the temporal lobe. Hum Brain Mapp 2003;19(4):224–47. http://dx.doi.org/10.1002/hbm.10123.

[41] Wen J, et al. Convolutional neural networks for classification of Alzheimer's disease: Overview and reproducible evaluation. Med Image Anal 2020;63:101694. http://dx.doi.org/10.1016/j.media.2020.101694.

[42] Qureshi MNI, Oh J, Lee B. 3D-CNN based discrimination of schizophrenia using resting-state fMRI. Artif Intell Med 2019;98:10–7. http://dx.doi.org/10.1016/j.artmed.2019.06.003.

[43] Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. In: Pereira F, Burges CJC, Bottou L, Weinberger KQ, editors. Adv Neural Inf Process Syst 2012;25. URL https://proceedings.neurips.cc/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf.

[44] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. In: Bengio Y, LeCun Y, editors. 3rd international conference on learning representations, ICLR 2015, san diego, CA, USA, May 7-9, 2015, conference track proceedings. 2015, URL http://arxiv.org/abs/1409.1556.

[45] Ioffe S, Szegedy C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: Proceedings of the 32nd international conference on international conference on machine learning - volume 37. ICML '15, JMLR.org; 2015, p. 448–56.

[46] Clevert D, Unterthiner T, Hochreiter S. Fast and accurate deep network learning by exponential linear units (ELUs). In: Bengio Y, LeCun Y, editors. 4th international conference on learning representations, ICLR 2016, san juan, puerto rico, May 2-4, 2016, conference track proceedings. 2016, URL http://arxiv.org/abs/1511.07289.

[47] Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: A simple way to prevent neural networks from overfitting. J Mach Learn Res 2014;15(56):1929–58, URL http://jmlr.org/papers/v15/srivastava14a.html.

[48] Chollet F, et al. Keras. 2015, https://keras.io.

[49] Abadi M, Agarwal A, Barham P, Brevdo E, Chen Z, Citro C, Corrado GS, Davis A, Dean J, Devin M, Ghemawat S, Goodfellow I, Harp A, Irving G, Isard M, Jia Y, Jozefowicz R, Kaiser L, Kudlur M, Levenberg J, Mané D, Monga R, Moore S, Murray D, Olah C, Schuster M, Shlens J, Steiner B, Sutskever I, Talwar K, Tucker P, Vanhoucke V, Vasudevan V, Viégas F, Vinyals O, Warden P, Wattenberg M, Wicke M, Yu Y, Zheng X. TensorFlow: Large-scale machine learning on heterogeneous systems. 2015, URL https://www.tensorflow.org/, Software available from tensorflow.org.

[50] Kingma DP, Ba J. Adam: A method for stochastic optimization. In: Bengio Y, LeCun Y, editors. 3rd international conference on learning representations, ICLR 2015, san diego, CA, USA, May 7-9, 2015, conference track proceedings. 2015, URL http://arxiv.org/abs/1412.6980.

[51] Mattia GM. CNNDiscriminationAbility. 2023, https://github.com/GiuliaMariaMattia/CNNDiscriminationAbility.

[52] Liu Y, Yu Y, Ouyang J, Jiang B, Yang G, Ostmeier S, Wintermark M, Michel P, Liebeskind DS, Lansberg MG, Albers GW, Zaharchuk G. Functional outcome prediction in acute ischemic stroke using a fused imaging and clinical deep learning model. Stroke 2023;54:2316–27. http://dx.doi.org/10.1161/STROKEAHA.123.044072.

[53] Kazuhiro K, et al. Generative adversarial networks for the creation of realistic artificial brain magnetic resonance images. Tomography 2018;4:159–63. http://dx.doi.org/10.18383/j.tom.2018.00042.

[54] Laino ME, Cancian P, Politi LS, Della Porta MG, Saba L, Savevski V. Generative adversarial networks in brain imaging: A narrative review. J Imaging 2022;8(4). http://dx.doi.org/10.3390/jimaging8040083, URL https://www.mdpi.com/2313-433X/8/4/83.

[55] Baur C, Denner S, Wiestler B, Navab N, Albarqouni S. Autoencoders for unsupervised anomaly segmentation in brain MR images: A comparative study. Med Image Anal 2021;69. http://dx.doi.org/10.1016/j.media.2020.101952.

[56] Kascenas A, Pugeault N, O'Neil AQ. Denoising autoencoders for unsupervised anomaly detection in brain MRI. In: Konukoglu E, Menze B, Venkataraman A, Baumgartner C, Dou Q, Albarqouni S, editors. Proceedings of the 5th international conference on medical imaging with deep learning. Proceedings of machine learning research, Vol. 172, PMLR; 2022, p. 653–64, URL https://proceedings.mlr.press/v172/kascenas22a.html.

[57] Baur C, Wiestler B, Muehlau M, Zimmer C, Navab N, Albarqouni S. Modeling healthy anatomy with artificial intelligence for unsupervised anomaly detection in brain MRI. Radiol: Artif Intell 2021;3. http://dx.doi.org/10.1148/ryai.2021190169.

[58] Muñoz-Ramírez V, Kmetzsch V, Forbes F, Meoni S, Moro E, Dojat M. Subtle anomaly detection: Application to brain MRI analysis of de novo parkinsonian patients. Artif Intell Med 2022;125. http://dx.doi.org/10.1016/j.artmed.2022.102251.

[59] Pinaya WH, Tudosiu PD, Gray R, Rees G, Nachev P, Ourselin S, Cardoso MJ. Unsupervised brain imaging 3D anomaly detection and segmentation with transformers. Med Image Anal 2022;79. http://dx.doi.org/10.1016/j.media.2022.102475.

[60] Villain E, Mattia GM, Nemmi F, Péran P, Franceries X, Le Lann MV. Visual interpretation of CNN decision-making process using simulated brain MRI. In: 2021 IEEE 34th international symposium on computer-based medical systems. CBMS, 2021, p. 515–20. http://dx.doi.org/10.1109/CBMS52027.2021.00102.