Deep Learning in Brain MRI: Effect of Data Leakage Due to Slice-level Split Using 2D Convolutional Neural Networks

Ekin Yagis

University of Essex

Selamawet Workalemahu Atnafu

University of Bologna

Alba García Seco de Herrera

University of Essex

Chiara Marzi

University of Bologna

Marco Giannelli

Azienda Ospedaliero-Universitaria Pisana

Carlo Tessa

Azienda USL Toscana Nord Ovest

Luca Citi

University of Essex

Stefano Diciotti (✉ stefano.diciotti@unibo.it)

University of Bologna

Research Article

Keywords: convolutional neural networks, cross-validation, data leakage, deep learning, neuroimaging

Posted Date: April 28th, 2021

DOI: <https://doi.org/10.21203/rs.3.rs-464091/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License. [Read Full License](#)

Abstract

In recent years, 2D convolutional neural networks (CNNs) have been extensively used for the diagnosis of neurological diseases from magnetic resonance imaging (MRI) data due to their potential to discern subtle and intricate patterns. Despite the high performances reported in numerous studies, developing CNN models with good generalization abilities is still a challenging task due to possible data leakage introduced during cross-validation (CV). In this study, we quantitatively assessed the effect of a data leakage caused by 3D MRI data splitting based on a 2D slice-level using three 2D CNN models for the classification of patients with Alzheimer's disease (AD) and Parkinson's disease (PD). Our experiments showed that slice-level CV erroneously boosted the average slice level accuracy on the test set by 30% on Open Access Series of Imaging Studies (OASIS), 29% on Alzheimer's Disease Neuroimaging Initiative (ADNI), 48% on Parkinson's Progression Markers Initiative (PPMI) and 55% on a local de-novo PD Versilia dataset. Further tests on a randomly labeled OASIS-derived dataset produced about 96% of (erroneous) accuracy (slice-level split) and 50% accuracy (subject-level split), as expected from a randomized experiment. Overall, the extent of the effect of an erroneous slice-based CV is severe, especially for small datasets.

1. Introduction

Deep learning has become a popular class of machine learning algorithms in computer vision and has been successfully employed in various tasks, including multimedia analysis (image, video, and audio analysis), natural language processing, and robotics¹. In particular, deep convolutional neural networks (CNNs) hierarchically learn high-level and complex features from input data, hence eliminating the need for handcrafting features, as in the case of conventional machine learning schemes².

The application of these methods in neuroimaging is rapidly growing (see Greenspan et al.³ and Zaharchuk et al.⁴ for reviews). Several studies employed deep learning methods for image improvement and transformation^{5–10}. Other studies performed lesion detection and segmentation^{11–13} and image-based diagnosis using different CNNs architectures^{14,15}. Deep learning has also been applied to more complex tasks, including identifying patterns of disease subtypes, determining risk factors, and predicting disease progression (see, e.g., Zaharchuk et al.⁴ and Davatzikos¹⁶ for reviews). Early works applied stacked auto encoders^{14,17,18} and deep belief networks¹⁹ to classify neurological patients from healthy subjects using data collected from different neuroimaging modalities, including magnetic resonance imaging (MRI), positron emission tomography (PET), resting-state functional MRI (rsfMRI), and the combination of these modalities²⁰.

Some authors reported very high accuracies in classifying patients with neurological diseases, such as Alzheimer's disease (AD) and Parkinson's disease (PD). For a binary classification of AD vs. healthy controls, Hon and Khan²¹ reported accuracy up to 96.25% using a transfer learning strategy. Sarraf et al.²² classified subjects as AD or healthy controls with a subject-level accuracy of 100% by adopting LeNet-5 and GoogleNet network architectures. In other studies, CNNs have been used for performing multi-class discrimination of subjects. Recently, Wu and colleagues²³ adopted a pre-trained CaffeNet and achieved accuracies of 98.71%, 72.04%, and 92.35% for a three-way classification between healthy controls, stable mild cognitive impairment (MCI), and progressive MCI patients, respectively. In another work by Islam and Zhang²⁴, an ensemble system of three homogeneous CNNs has been proposed and average multi-class classification accuracy of 93.18% was found on the Open Access Series of Imaging Studies (OASIS) dataset. For the classification of PD, Esmailzadeh et al.²⁵ classified PD patients from healthy controls based on MRI and demographic information (i.e., age and gender). With the proposed 3D model,

they achieved 100% accuracy on the test set. In another study by Sivaranjini and Sujatha²⁶, a pre-trained 2D CNN AlexNet architecture was used to classify PD patients vs. healthy controls, resulting in an accuracy of 88.9%.

Although very good performances have been shown by using deep learning for the classification of neurological disorders, there are still many challenges that need to be addressed, including complexity and difficulty in interpreting the results due to highly nonlinear computations, non-reproducibility of the results, and data/information and, especially, data overfitting (see Vieira et al.²⁰ and Davatzikos¹⁶ for reviews).

Overly optimistic results may be due to data leakage – a process caused by incorporating information of test data into the learning process. While concluding that data leakage leads to overly optimistic results will surprise few practitioners, we believe that the extent to which this is happening in neuroimaging applications is mostly unknown, especially in small datasets. As we completed this study, we became aware of independent research by Wen et al.²⁷ that corroborates part of our conclusions regarding the problem of data leakage. They successfully suggested a framework for reproducible assessment of AD classification methods. However, the architectures have not been trained and tested on smaller datasets typical of clinical practice and they mainly employed hold-out model validation strategies, rather than cross-validation (CV) – that gives a better indication of how well a model performs on unseen data^{28,29}. Moreover, the authors focused on illustrating the effect of data leakage on the classification of AD patients only.

Unfortunately, the problem of data leakage incurred by incorrect data split is not only limited within the area of AD classification but can also be seen in various neurological disorders. It is more common to observe the data leakage in 2D architectures, yet some forms of data leakage, such as late split, could be present in 3D CNN studies as well. Moreover, although deep complex classifiers are more prone to overfitting, there is no reason to believe that conventional machine learning algorithms could not be affected by data leakage. A summary of these works with clear and potential data leakage is given in Tables 1 and 2, respectively. Other works with insufficient information to assess data leakage are reported in Table 3.

Table 1

Summary of the previous studies performing classification of neurological disorders using MRI and with clear data leakage.

Disorder	Reference	Groups (number of subjects)	Machine learning model	Data split method	Type of data leakage	Accuracy (%)
	Gunawardena et al., 2017 ⁵⁸	AD-MCI-HC (36)	2D CNN	4:1 train/test slice-level split	wrong split	96.00
	Hon & Khan, 2017 ²¹	AD-HC (200)	2D CNN (VGG16)	4:1 train/test slice-level split	wrong split	96.25
	Jain et al., 2019 ⁵⁹	AD-MCI-HC (150)	2D CNN (VGG16)	Data augmentation + 4:1 train/test slice-level split	late and wrong split	95.00
	Khagi et al., 2019 ⁶⁰	AD-HC (56)	2D CNN (AlexNet, GoogLeNet, ResNet50, new CNN)	6:2:2 train/validation/test slice-level split	wrong split	98.00
AD/MCI	Sarraf et al., 2017 ²²	AD-HC (43)	2D CNN (LeNet-5)	3:1:1 train/validation/test slice-level split	wrong split	96.85
	Wang et al., 2017 ⁶¹	MCI-HC (629)	3D CNN	Data augmentation + 10:3:3 train/validation/test split by MRI slices	wrong split	90.60
	Puranik et al., 2018 ⁶²	AD/EMCI-HC	2D CNN	17:3 train/test split by MRI slices	wrong split	98.40
	Basheera et al., 2019 ⁶³	AD-HC	2D CNN	4:1 train/test split by MRI slices	wrong split	90.47
	Basaia et al., 2019 ⁶⁴	AD-HC	2D CNN	9:1 slice level split	wrong split	99.00

AD = Alzheimer's disease; HC = Healthy controls; MCI = Mild cognitive impairment.

Table 2

Summary of the previous studies performing classification of neurological disorders using MRI and suspected to have potential data leakage.

Disorder	Reference	Groups (number of subjects)	Machine learning model	Data split method	Type of data leakage	Accuracy (%)
	Farooq et al., 2017 ³⁵	AD-MCI-LMCI-HC (355)	2D CNN (GoogLeNet and modified ResNet)	3:1 train/test (potential) slice-level split	wrong split	98.80
AD/MCI	Ramzan et al., 2019 ⁶⁵	HC-SMC-EMCI-MCI-LMCI-AD (138)	2D CNN (ResNet-18)	7:2:1 train/validation/test (potential) slice-level split	wrong split	100
	Raza et al., 2019 ⁶⁶	AD-HC (432)	2D CNN (AlexNet)	4:1 train/test (potential) slice-level split	wrong split	98.74
	Pathak et al., 2020 ⁶⁷	AD-HC	2D CNN	3:1 (potential) slice level split	wrong split	91.75
ASD	Libero et al., 2015 ⁶⁸	ASD-TD (37)	Decision tree	unclear	entire data set used for feature selection	91.90
	Zhou et al., 2014 ⁶⁹	ASD-TD/HC (280)	Random tree classifier	4:1 train/test split	entire data set used for feature selection	100
PD	Sivaranjini, et al., 2019 ²⁶	PD-HC (182)	2D CNN	4:1 train/test split by MRI slices	wrong split	88.90
TBI	Lui et al., 2014 ⁷⁰	TBI-HC (47)	Multilayer perceptron	10-fold CV	entire data set used for feature selection	86.00
Brain tumor	Hasan et al., 2019 ⁷¹	Tumor-HC (600)	MGLCM + 2D CNN + SVM	10-fold CV	dataset is divided randomly into 0 folds	99.30

AD = Alzheimer's disease; ASD = Autism spectrum disorder; HC = Healthy controls; MCI = Mild cognitive impairment; PD = Parkinson's disease; SWEDD = scans without evidence of dopaminergic deficit; TBI = Traumatic brain injury; TD = Typically developing.

Table 3

Summary of the previous studies performing classification of neurological disorders using MRI and that provide insufficient information to assess data leakage.

Disorder	Reference	Groups (number of subjects)	Machine learning model	Data split method	Accuracy (%)
	Al-Khuzai et al., 2021 ⁷²	AD-HC (240)	2D CNN	(potential) slice-level split	99.30
AD/MCI	Billones et al., 2016 ⁷³	AD-MCI-HC (900)	2D CNN (modified VGG16)	7:3 train/test (potential) slice-level split	91.85
	Wu et al., 2018 ²³	AD-HC	2D CNN	Data augmentation + 2:1 train/test split by MRI slices	97.58

AD = Alzheimer's disease; HC = Healthy controls; MCI = Mild cognitive impairment.

In this study, we addressed the issue of data leakage in one of the most common classes of deep learning models, i.e., 2D CNNs, caused by incorrect dataset split of 3D MRI data. Specifically, we quantified the effect of data leakage on CNN models trained on different datasets of T_1 -weighted brain MRI of healthy controls and patients with neurological disorders using a nested CV scheme with two different data split strategies: a) subject-level split, avoiding any form of data leakage and b) slice-level split, in which different slices of the same subject are contained both in the training and the test folds (thus data leakage will occur). We focused our attention on both large (about 200 subjects) and small (about 30 subjects) datasets to evaluate a possible increase of performance overestimation when a smaller dataset was used, as is often the case in clinical practice. This paper expands on the preliminary results by Yagis et al.³⁰ offering a broader investigation on the issue. In particular, we performed the classification of AD patients using the following datasets: 1) OASIS-200, consisting of randomly sampled 100 AD patients and 100 healthy controls from the OASIS-1 study³¹, 2) ADNI, including 100 AD patients and 100 healthy controls randomly sampled from Alzheimer's Disease Neuroimaging Initiative (ADNI)³², and 3) OASIS-34, composed of 34 subjects (17 AD patients and 17 healthy controls) randomly selected from the OASIS-200 dataset. Given that the performance of a model trained on a small sample dataset could depend on the selected samples, we created ten instances of the OASIS-34 dataset by randomly sampling from the OASIS-200 dataset ten times independently. The subject IDs included in each instance are found in Supplementary Table S1 (Supporting Information). Moreover, we generated a further dataset, called OASIS-random, where, for each subject of the OASIS-200 dataset, a fake random label of either AD patient or healthy control was assigned. In this case, the image data had no relationship with the assigned labels. Besides, we included two T_1 -weighted images datasets of patients with de-novo PD: PPMI, including 100 de-novo PD patients and 100 healthy controls randomly chosen from the public Parkinson's Progression Markers Initiative (PPMI) dataset³³, and Versilia, a small-sized private clinical dataset of 17 patients with de-novo PD and 17 healthy controls. A detailed description of each dataset has been reported in the "Methods" section.

2 Results

For AD classification, accuracies on the test set, using subject-level CV, were below 71% for large datasets (OASIS-200 and ADNI), whereas they were below 59% for smaller datasets (OASIS-34). As regards de novo PD classification, they were around 50% for both large (PPMI) and small (Versilia) datasets. Conversely, in all datasets, slice-level CV erroneously produced very high classification accuracies on the test set (higher than 94% and 92% on large and small datasets, respectively), leading to deceptive, over-optimistic results (Table 4).

The worst-case stemmed from the randomly labeled OASIS dataset, which resulted in a model with unacceptably high performances (accuracy on the test set more than 93%) using slice-level CV, whereas classification results obtained using a subject-level CV were about 50%, in accordance with the expected outcomes for a balanced dataset with completely random labels.

Table 4

Mean slice-level accuracy on the training and test set of the outer CV over 5-fold nested CV has been reported for three 2D CNN models (see “Methods” section), all datasets and two data split methods (slice-level and subject-level). The difference between accuracy using slice-level and subject-level split in the test set has also been reported.

Dataset	Network architecture	Training set accuracy (%)		Test set accuracy (%)		
		Subject-level split	Slice-level split	Subject-level split	Slice-level split	Difference
	VGG16-v1	95.93	99.85	66.0	94.18	28.18
OASIS-200	VGG16-v2	95.13	100	66.13	96.99	30.86
	ResNet-18	100	100	68.87	98.96	30.1
	VGG16-v1	88.94	100	54.35	99.19	44.84
OASIS-34	VGG16-v2	96.94	100	54.34	99.33	44.99
	ResNet-18	100	100	57.49	98.96	41.47
	VGG16-v1	63.38	100	53.37	95.93	42.56
OASIS-random	VGG16-v2	69.17	100	49.25	94.81	45.56
	ResNet-18	84.49	99.09	50.8	93.74	42.94
ADNI	VGG16-v1	91.09	100	70.12	95.31	25.19
	VGG16-v2	80.49	100	66.49	95.24	28.75
	ResNet-18	100	100	68.68	96.87	30.19
PPMI	VGG16-v1	76.8	100	48.24	93.99	45.75
	VGG16-v2	73.19	100	46.93	94.37	47.44
	ResNet-18	100	100	48.06	96.12	44.06
Versilia	VGG16-v1	99.72	100	53.86	95.97	42.11
	VGG16-v2	76.89	100	42.97	97.8	54.83
	ResNet-18	99.90	95.13	51.36	92.63	41.27

3 Discussion

In this study, we quantitatively assessed the extent of the overestimation of the model’s classification performance caused by the adoption of an incorrect slice-level CV, which is unfortunately frequent in neuroimaging literature (see Tables 1–3). More specifically, we showed the performance of three 2D CNN models (two VGG variants and one ResNet-18, see “Methods” section) trained with subject-level and slice-level CV data splits for the classification of AD and PD patients from healthy controls using T_1 -weighted brain MRI data. Our results revealed that pooling slices of MRI volumes for all subjects and then dividing randomly into training and test set leads to significantly inflated accuracies (in some cases from barely above chance level to about 99%). In particular, slice-level CV

erroneously increased the average slice level accuracy on the test set by 40–55% on smaller datasets (OASIS-34 and Versilia) and 25–45% on larger datasets (OASIS-200, ADNI, PPMI). Moreover, we also conducted an additional experiment in which all the labels of the subjects were fully randomized (OASIS-random dataset). Even under such circumstances, using the slice-level split, we achieved an erroneous 95% classification accuracy on the test set with all models, whereas we found 50% accuracy using a subject-level data split, as expected from a randomized experiment. This large (and erroneous) increase in performance could be due to the high intra-subject correlation among T_1 -weighted slices, resulting in a similar information content present in slices of the same subject³⁴.

In AD classification, three previous studies^{21,22,35}, using similar deep networks (VGG16, ResNet-18 and LeNet-5, respectively), reported higher classification accuracies (92.3%, 98.0% and 96.8%, respectively) than ours. However, there is a strong indication that these performances are massively overestimated due to a slice-level split. In particular, in one of these works²¹, the presence of data leakage was further corroborated by the source code accompanying the paper and confirmed by our data. In fact, when we used the same dataset of Hon and Khan²¹ (OASIS-200 dataset), our VGG16 models achieved only 66% classification accuracy with subject-level split, whereas they boosted to about 97% with a slice-level split. Similar findings were presented by Wen et al.²⁷ who used an ADNI dataset with 330 healthy controls and 336 AD patients. Indeed, using baseline data, they reported a 79% of balanced accuracy in the validation set with a subject-level split which increased up to 100% with a slice-level split.

One of the main issues in the classification of neurological disorders using deep learning is data scarcity³⁶. Not only because labeling is expensive, but also because privacy reasons and institutional policies make acquiring large sets of labeled imaging data even more challenging³⁷. To show the impact of data size on model performance, we created 10 small subsets from the OASIS dataset (OASIS-34 datasets). As expected, when we reduced the data, we obtained lower classification accuracies with all the networks using the subject-level data split method. However, when the slice-level method was used, the models achieved erroneous better results on OASIS-34 than on the OASIS-200 dataset. Similarly, models trained on the Versilia dataset (34 subjects) produced inflated results with the slice-level split. Overall, these results point out that the data leakage is extremely relevant especially when small datasets are used – a situation which may be unfortunately common in clinical practice.

It is well-known that data leakage leads to inflating performance. Nevertheless, the degree of overestimation quantified through our experiments was surprising. Unfortunately, in the literature, the precise application of CV is frequently not well-documented and the source code is not available. This situation leaves the neuroimaging community unable to trust the (sometimes) promising results published. Regardless of the network architecture, the number of subjects, and the level of complexity of the classification problem, all experiments that applied slice-level CV yielded very high classification accuracies on the test set as a result of incorporating different slices of the same subject in both the training and test sets. Considering classifications on 2D MRI images, to prevent data leakage and to get trustable results, we showed that it is crucial that the CV split be done based on the subject-level. This assures the training and validation sets to be completely independent and confirms that no information is leaking from the test set into the training set during the development of the model. With recent advances in machine learning, more and more people are becoming interested in applying these techniques to biomedical imaging, and there is a real and growing risk that many of them will not be familiar with the possible issues and good practices. We also emphasize the need for documenting how the CV is implemented, the architecture used, how the different hyperparameter choices/tunings are made, and also include their values where possible. Besides, it would be also necessary to make the source codes available to the neuroimaging community so that the results will be reproducible³⁸.

The main limitation of this study is that we have not assessed all data leakage types, including late split and hyperparameters optimization in the test set – that may be also both present in 3D CNN studies. A late split occurs when the data augmentation step is performed before separating the test set from the training data. In that case, the augmented data generated from the same original image can be seen in both training and test data, leading to inflated performance²⁷. Still, using the same test set for optimizing the training hyperparameters as well as evaluating the model performance is an additional form of data leakage³⁹. Finally, data leakage also occurs when feature selection is performed based on the whole dataset before carrying out cross-validation^{39,40}. We have found evidence of all these data leakage issues in the recent literature (see Tables 1–3) and we plan to systematically quantify their effect in our future work. Another limitation may be due to the substantial overfitting we observed. Indeed, using more efficient models, by searching for best architectures and exploring a wider hyperparameter space, better performances in the test set would have been achieved, and hence the boost in performance due to the slice-level split would be likely reduced.

In conclusion, training a 2D CNN model for analyzing 3D brain image data should be performed using a subject-level CV to prevent data leakage. The adoption of slice-based CV results in very optimistic model performances, especially for small datasets, as the extent of the overestimation due to data leakage is severe.

4 Materials And Methods

4.1 Datasets

In this study, we adopted the scans collected by three public and international datasets of T₁-weighted images of patients with AD (the OASIS dataset³¹ and the ADNI dataset³²) and de-novo PD (the PPMI dataset³³). An additional private de-novo PD dataset, namely the Versilia dataset, has also been used. A summary of the demographics of the datasets used in this study is shown in Table 5. In the following sections, a detailed description of all datasets will be reported.

Table 5

Demographic features of subjects belonging to OASIS-200, ADNI, PPMI and Versilia datasets. The same information for the OASIS-34 datasets have been reported in Supplementary Table S5 (Supporting Information).

Dataset		Patients	Healthy controls
OASIS-200	Number of subjects	100	100
	Age (range, years)	62–96	59–94
	Age (mean $\pm$ SD, years)	76.70 $\pm$ 7.10	75.50 $\pm$ 9.10
	Gender (women/men)	59/41	73/27
ADNI	Number of subjects	100	100
	Age (range, years)	56–89	58–95
	Age (mean $\pm$ SD, years)	74.28 $\pm$ 7.96	75.04 $\pm$ 7.11
	Gender (women/men)	44/56	52/48
PPMI	Number of subjects	100	100
	Age (range, years)	34–82	31–83
	Age (mean $\pm$ SD, years)	61.71 $\pm$ 9.99	61.91 $\pm$ 11.52
	Gender (women/men)	40/60	36/64
Versilia	Number of subjects	17	17
	Age (range, years)	48–78	54–77
	Age (mean $\pm$ SD, years)	64 $\pm$ 7.21	64.00 $\pm$ 7.00
	Gender (women/men)	4/13	5/12

AD = Alzheimer's disease; ADNI = Alzheimer's Disease Neuroimaging Initiative; OASIS = Open Access Series of Imaging Studies; PD = Parkinson's disease; PPMI = Parkinson's Progression Markers Initiative; SD = standard deviation.

4.1.1 OASIS-200, OASIS-34, and OASIS-random datasets

We have used the T_1 -weighted images of 100 AD patients [(59 women and 41 men, age 76.70 $\pm$ 7.10 years, mean $\pm$ standard deviation (SD))] and 100 healthy controls (73 women and 27 men, age 75.50 $\pm$ 9.10 years, mean $\pm$ SD) from the OASIS-1 study – a cross-sectional cohort of the OASIS brain MRI dataset³¹, freely available at <https://www.oasis-brains.org/>. In particular, we have employed the same scans that were previously selected by other authors²¹. We called this dataset OASIS-200. The subject identification numbers (IDs) and demographics of these subjects were specified in Supplementary Table S2 (Supporting Information). No significant difference in age ($p = 0.15$ at t-test) was found between the two groups, while a significant (borderline) difference in gender was observed ($p = 0.04$ at χ^2 -test).

In OASIS-1, AD diagnosis, as well as the severity of the disease, were evaluated based on the global Clinical Dementia Rating (CDR) score derived from individual CDR scores for the domains memory, orientation, judgment

and problem solving, function in community affairs, home and hobbies, and personal case^{41,42}. Subjects with a global CDR score of 0 have been labeled as healthy controls, while scores 0.5 (very mild), 1 (mild), 2 (moderate), and 3 (severe) have been all labeled as AD.

All T_1 -weighted images have been acquired on a 1.5 T MR scanner (Vision, Siemens, Erlangen, Germany), using a Magnetization Prepared Rapid Gradient Echo (MPRAGE) sequence in a sagittal plane [repetition time (TR) = 9.7 ms, echo time (TE) = 4.0 ms, flip angle = 10°, inversion time (TI) = 20 ms, delay time (TD) = 200 ms, voxel size = 1 mm × 1 mm × 1.25 mm, matrix size = 256 × 256, number of slices = 128]³¹.

4.1.2 ADNI dataset

We considered the T_1 -weighted MRI data of 100 AD patients (44 women and 56 men, age 74.28 ± 7.96 years, mean $\pm$ SD) and 100 healthy controls (52 women and 48 men, age 75.04 ± 7.11 years, mean $\pm$ SD). No significant difference in age ($p = 0.24$ at t-test) and gender ($p = 0.26$ at χ^2 -test) was found between the two groups. Alzheimer's disease patients have been randomly chosen from the ADNI 2 dataset (available at <http://adni.loni.usc.edu/>) – a cohort of ADNI that extends the work of ADNI 1 and ADNI-GO studies³². Led by the Principal Investigator Michael W. Weiner, MD, ADNI was launched in 2003 with the aim of investigating if biological markers (such as MRI and PET) can be combined to define the progression of MCI and early AD. We have used MPRAGE T_1 -weighted MRI scans acquired by 3 T scanners [6 Siemens (Erlangen, Germany) MRI scanners and 6 Philips (Amsterdam, Netherlands) scanners] in a sagittal plane (voxel size = 1 mm × 1 mm × 1.2 mm). The image size of the T_1 -weighted data acquired from the Siemens and Philips scanners were $176 \times 240 \times 256$ and $170 \times 256 \times 256$, respectively. Since ADNI 2 is a longitudinal dataset, more than one scan was available for each subject. The first scan of each participant has been chosen to produce a cross-sectional dataset. Supplementary Table S3 (Supporting Information) provides subject IDs and the acquisition date of the specific scan used in our study. The MRI acquisition protocol for each MRI scanner can be found at <http://adni.loni.usc.edu/methods/documents/mri-protocols/>. In ADNI 2 dataset, subjects have been categorized as AD patients or healthy controls based on whether subjects have complaints about their memory and by considering a combination of neuropsychological clinical scores³².

4.1.3 PPMI dataset

We randomly selected 100 de-novo PD subjects (40 women and 60 men, age 61.71 ± 9.99 , mean $\pm$ SD) and 100 healthy controls (36 women and 64 men, age 61.91 ± 11.52 , mean $\pm$ SD) from the publicly available PPMI dataset (<https://ida.loni.usc.edu/login.jsp?project=PPMI>). No significant difference in age ($p = 0.44$ at t-test) and gender ($p = 0.56$ at χ^2 -test) was found between the two groups. The criterion used to recruit de-novo PD patients, and healthy controls were defined by Marek and colleagues³³. Briefly, PD patients were selected within two years of diagnosis with a Hoehn and Yahr score < 3 ⁴³, at least two of resting tremor, either bradykinesia or rigidity (must have either resting tremor or asymmetric bradykinesia) or a single asymmetric resting tremor or asymmetric bradykinesia and dopamine transporter (DAT) or vesicular monoamine transporter type 2 (VMAT-2) imaging showing a dopaminergic deficit. Healthy controls were free from any clinically significant neurological disorder³³.

The T_1 -weighted scans were collected at baseline using MR scanners manufactured by Siemens (11 scanners at 3 T and 5 scanners at 1.5 T), Philips Medical Systems (10 scanners at 3 T and 11 scanners at 1.5 T), GE Medical Systems (11 scanners at 3 T and 24 scanners at 1.5 T) and another anonymous one (5 scanners at 1.5 T). We also

found three subjects whose MRI protocol was missing. The details of the MRI protocols of all scanners can be found in Supplementary Table S4 (Supporting Information).

4.1.4 Versilia dataset

Seventeen (4 women and 13 men, age 64 ± 7.21 years, mean $\pm$ SD) patients with de-novo parkinsonian syndrome consecutively referred to a Neurology Unit for the evaluation of PD over a 24-month interval (from June 2012 to June 2014) were recruited in this dataset. More details about clinical evaluation can be found in⁴⁴. Seventeen healthy controls (5 women and 12 men, age 64 ± 7 years, mean $\pm$ SD) with no history of neurological diseases and normal neurological examination were recruited as controls. No significant difference in age ($p = 0.95$ at t-test) and gender ($p = 0.70$ at χ^2 -test) was found between the two groups.

All subjects underwent high-resolution 3D T_1 -weighted imaging on a 1.5 T MR scanner system (Magnetom Avanto, software version Syngo MR B17, Siemens, Erlangen-Germany) equipped with a 12-element matrix radiofrequency head coil and SQ-engine gradients. The SQ-engine gradients had a maximum strength of 45 mT/m and a slew rate of 200 T/m/s. T_1 -weighted MR images were acquired with an axial high resolution 3D MPRAGE sequence with TR = 1900 ms, TE = 3.44 ms, TI = 1100 ms, flip angle = 15° , slice thickness = 0.86 mm, field of view (FOV) = 220 mm $\times$ 220 mm, matrix size = 256 $\times$ 256, number of excitations (NEX) = 2, number of slices = 176.

4.2 T_1 -weighted MRI data pre-processing

All T_1 -weighted MRI data went through two preprocessing steps (see Fig. 1). In the first stage, co-registration to a standard template space and skull stripping were applied to geometrically re-align all the images and remove non-brain regions. In the second stage, the collection of a subset of axial images using an entropy-based slice selection approach has been carried out.

4.2.1 Co-registration to a standard template space and skull stripping

For the OASIS datasets, we used publicly available pre-processed data (gain-field corrected, brain masked, and co-registration)⁴⁵. Briefly, the brain masks from OASIS were obtained using an atlas-registration-based method, and their quality was controlled by human experts³¹ and each volume has been co-registered to the Talairach and Tournoux atlas. Each pre-processed T_1 -weighted volume had a data matrix size of 176 $\times$ 208 $\times$ 176 and a voxel size of 1 mm $\times$ 1 mm $\times$ 1 mm⁴⁵.

For all other datasets, we have co-registered each individual T_1 -weighted volume to the MNI152 standard template space (at 1 mm voxel size – available in the FSL version 6.0.3 package) by using the SyN algorithm included in ANTs package (version 2.1.0) with default parameters⁴⁶. Then, the brain mask of the standard template space has been applied to each co-registered volume. Each pre-processed T_1 -weighted volume had a data matrix size of 182 $\times$ 218 $\times$ 182 and a voxel size of 1 mm $\times$ 1 mm $\times$ 1 mm.

Supplementary Fig. S1 (Supporting Information) illustrates sample pre-processed T_1 -weighted slices from OASIS-200, ADNI, PPMI, and Versilia datasets.

4.2.2 Entropy-based slice selection

Each T_1 -weighted slice generally conveys a different amount of information. Given that we are interested in developing a 2D CNN model, we have performed a preliminary slice selection based on the amount of information, retaining, for each T_1 -weighted volume, only the eight axial slices that showed the highest entropy²¹. Specifically, for a slice with k grayscale levels and with each gray level having a probability of occurrence p_k (estimated as its relative frequency in the image), the Shannon entropy E_S was computed as:

$$E_S = -\sum_k p_k \log_2(p_k) \quad (1)$$

To be consistent with the input sizes of the proposed 2D CNN models, all slices were resized to 224×224 pixels by fitting a cubic spline between the 4-by-4 neighborhood pixels⁴⁷. Voxel-wise feature standardization has also been applied to make training the CNNs easier and to achieve faster convergence, i.e., for each voxel, an average value of all grayscale values within the brain mask has been subtracted and scaled by the standard deviation (within the brain mask).

4.3 Model architectures

Since the number of subjects of each dataset may not be sufficient to train with high accuracy a 2D CNN model from scratch, we have used a machine learning technique called transfer learning that allows employing pre-trained models, i.e., model parameters previously developed for one task (source domain) to be transferred to target domain for weight initialization and feature extraction. In particular, CNN layers hierarchically extract features starting from the general low-level features to those specific to the target class, and, using transfer learning, the general low-level features can be shared across tasks. Notably, in this study, we used pre-trained VGG16⁴⁸ and ResNet-18⁴⁹ models, as detailed in the following sections. The transfer learning approach and VGG16 architectures used in this study are similar to those employed in²¹ as their results triggered our investigation of data leakage.

4.3.1 VGG16-based models

VGG16 is one of the most influential architectures which explores network depth with very small (3x3) convolution filters stacked on top of each other. VGG16 consists of five convolutional blocks, with alternating convolutional and pooling layers, and three fully-connected layers.

In transfer learning, the most common approach is copying the first n layers of the pre-trained network to the first n layers of a target network, and then randomly initializing the remaining layers to be trained on the target task. Depending on the size of the target dataset and the number of parameters in the first n layers, these copied features can be left unchanged (i.e., frozen) or fine-tuned during the training of the network on a new dataset. It is well accepted that if the target dataset is relatively small, fine-tuning may cause overfitting, whereas if the target dataset is large, then the base features can be fine-tuned to improve the performance of the model without overfitting.

To investigate the effect of fine-tuning, we have tested two different variants of VGG16 architecture, namely VGG16-v1 and VGG16-v2 (Fig. 2). The former model has been used as a feature extractor where the weights for all layers of the network are frozen except that of the final fully connected layer. The three topmost layers have been replaced by randomly initialized fully connected layers with rectified linear unit (ReLU) activation. The weights are initialized according to the Xavier initialization heuristic⁵⁰ to prevent the gradients from vanishing or exploding.

The VGG16-v2 model has been utilized as a weight initializer where the weights are derived from the pre-trained network and fine-tuned during training. We have replaced the fully connected layers with a randomly initialized global average pooling (GAP) layer suggested by Lin and colleagues⁵¹ to reduce the number of parameters and, rather than freezing the CNN layers, we have fine-tuned all layers.

4.3.2 ResNet-18 based model

It has been long believed that deeper networks can learn more complex non-linear relationships than shallower networks with the same number of neurons, and thus network depth is of great importance on model performance⁵². However, many studies revealed that deeper networks often converge at a higher training and test error rate when compared to their shallower counterparts⁴⁹. Therefore, stacking more layers to the plain networks may eventually degrade the model's performance while complicating the optimization process. To overcome this issue, He and colleagues introduced deep residual neural networks and achieved top-5 test accuracies with their models on the popular ImageNet test set⁴⁹. The model was proposed as an attempt to solve the vanishing gradients and the degradation problems using residual blocks. With these residual blocks, the feature of any deeper unit can be computed as the sum of the activation of a shallower unit and the residual function. This architecture causes the gradient to be directly propagated to shallower units making ResNets easier to train.

There are different versions of ResNet architecture with various numbers of layers. In this work, we used ResNet-18 architecture, which is an 18-layer residual deep learning network consisting of five stages, each with a convolution and identity block⁴⁹. In our model, one fully connected layer with sigmoid activation has been added at the end of the network – a common practice in binary classification tasks as it takes a real-valued input and squashes the output to a range between 0 and 1. Since the network is relatively smaller and has a lower number of parameters than VGG16, the weights and biases of all the transferred layers are fine-tuned while the newly added fully connected layer has been trained starting from randomly initialized weights. The architecture of our ResNet-18 model can be seen in Fig. 3.

4.3 Model training and validation

Each 2D CNN model has been trained and validated using a nested CV strategy – a validation scheme that allows examining the unbiased generalization performance of the trained models along with performing, at the same time, hyperparameters optimization³⁹. It involves nesting two k -fold CV loops where the inner loop is used for optimizing model hyperparameters, and the outer loop gives an unbiased estimate of the performance of the best model. It is especially suitable when the amount of data available is insufficient to allow separate validation and test sets³⁹. A schematic diagram of the procedure is illustrated in Supplementary Fig. S2 (Supporting Information). It starts by dividing the dataset into k folds and one-fold is kept as a test set (outer CV), while the other $k-1$ folds are split into inner folds (inner CV). The model hyperparameters are chosen from the hyperparameter space through a grid search based on the average performance of the model over the inner folds. In particular, we varied the learning rate in the set $\{10^{-5}, 3 \times 10^{-5}, 10^{-4}, 3 \times 10^{-4}, 10^{-3}\}$ and the learning rate decay in $\{0, 0.1, 0.3, 0.5\}$. The chosen model is then fitted with all the outer fold training data and tested on the unseen test fold, resulting in an unbiased estimation of the model's prediction error. Specifically, we choose a 10-fold CV because it offers a favorable bias-variance tradeoff^{53,54}.

In all experiments, we used batch size = 128 and epoch number = 50. Due to its ability to adaptively updating individual learning rates for each parameter, an Adam optimizer was used⁵⁵. Each selected slice of the 3D T_1 -

weighted volume has been classified independently and the final model's performance was stated using the mean slice-level accuracy, separately, on the training set and test set folds of the outer CV.

We thus conducted CNNs model's training and validation on each dataset in a nested CV loop using two different data split strategies: a) subject-level split, in which all the slices of a subject have been placed either in the training set or in the test set, avoiding any form of data leakage; b) slice-level split, in which all the slices have been pooled together before CV, then split randomly into training and test set. In this case, for each slice of the test set, a set of highly correlated slices coming from the MR volume of the same subject ended up in the training set, giving rise to data leakage, as shown pictographically in Fig. 1.

CNN models were carried out using a custom-made software in Python language (version 3.6.8) using the following modules: CUDA v.9.0.176, TensorFlow-gpu v.1.12.0, Keras v.2.2.4⁵⁶, Scikit-learn v.0.20.2⁵⁷, Nibabel v.2.3.3, and OpenCV v.3.3.0. All the code, model architectures, and model weights will be made publicly available at the time of acceptance of this manuscript. The training and validation of CNN models were performed on a workstation equipped with a 12 GB G5X frame buffer NVIDIA TITAN X (Pascal) GPU with 64 GB RAM, 8 CPUs, 3584 CUDA cores and 11.4 Gbps processing speed. The average computational time for CNN training on a dataset of 34 and 200 subjects were 5.68 hours (VGG16-v1), 5.63 hours (VGG16-v2), 2.94 hours (ResNet-18) and 33.93 hours (VGG16-v1), 33.82 hours (VGG16-v2), 14.12 hours (ResNet-18), respectively. The total computational time for this study was thus about 17 days.

Declarations

Acknowledgements

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

http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf

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

at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

Data were provided [in part] by OASIS: Cross-Sectional: Principal Investigators: D. Marcus, R. Buckner, J. Csernansky J. Morris; P50 AG05681, P01 AG03991, P01 AG026276, R01 AG021910, P20 MH071616, U24 RR021382.

Data used in the preparation of this article were [in part] obtained from the Parkinson's Progression Markers Initiative (PPMI) database www.ppmi-info.org/data. For up-to-date information on the study, visit www.ppmi-info.org.

PPMI – a [public-private partnership](#) – is funded by the Michael J. Fox Foundation for Parkinson's Research and multiple funding partners. The full list of PPMI funding partners can be found at ppmi-info.org/fundingpartners.

Funding

This work was supported in part by the European Union's Horizon 2020 research and innovation programme under Grant Agreement No 824153 “POTION” (EY and LC) and in part by an NVIDIA Academic GPU Grant Program (SD).

Author contributions

E.Y., S.W.A., A.G., L.C., and S.D. conceived the original idea. C.M. and S.D. performed the data pre-processing. E.Y. and S.W.A developed the software implementation of the model training and validation. E.Y., S.W.A, C.M., and S.D. drafted the manuscript and designed figures and tables. All authors contributed to the interpretation of the results and revised the final version of the manuscript.

Additional information

The authors declare no competing interests.

References

1. Hatcher, W. G. & Yu, W. A Survey of Deep Learning: Platforms, Applications and Emerging Research Trends. *IEEE Access* **6**, 24411–24432 (2018).
2. Goodfellow, I., Bengio, Y. & Courville, A. *Deep learning*. (The MIT Press, 2016).
3. Greenspan, H., van Ginneken, B. & Summers, R. M. Guest Editorial Deep Learning in Medical Imaging: Overview and Future Promise of an Exciting New Technique. *IEEE Trans. Med. Imaging* **35**, 1153–1159 (2016).
4. Zaharchuk, G., Gong, E., Wintermark, M., Rubin, D. & Langlotz, C. P. Deep Learning in Neuroradiology. *AJNR Am J Neuroradiol* **39**, 1776–1784 (2018).
5. Bahrami, K. *et al.* Reconstruction of 7T-Like Images From 3T MRI. *IEEE Trans Med Imaging* **35**, 2085–2097 (2016).
6. Han, X. MR-based synthetic CT generation using a deep convolutional neural network method. *Med. Phys.* **44**, 1408–1419 (2017).
7. Li, R. *et al.* Deep learning based imaging data completion for improved brain disease diagnosis. in *MICCAI International Conference on Medical Image Computing and Computer-Assisted Intervention* vol. 17 305–312 (2014).

8. Liu, F., Jang, H., Kijowski, R., Bradshaw, T. & McMillan, A. B. Deep Learning MR Imaging–based Attenuation Correction for PET/MR Imaging. *Radiology* **286**, 676–684 (2018).
9. Vemulapalli, R. Deep Networks and Mutual Information Maximization for Cross-Modal Medical Image Synthesis. 381–403 (2017) doi:10.1016/B978-0-12-810408-8.00022-5.
10. Zhu, B., Liu, J. Z., Cauley, S. F., Rosen, B. R. & Rosen, M. S. Image reconstruction by domain-transform manifold learning. *Nature* **555**, 487–492 (2018).
11. Chang, P. D. Fully Convolutional Deep Residual Neural Networks for Brain Tumor Segmentation. in *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries* (eds. Crimi, A. et al.) vol. 10154 108–118 (Springer International Publishing, 2016).
12. Dou, Q. *et al.* Automatic Detection of Cerebral Microbleeds From MR Images via 3D Convolutional Neural Networks. *IEEE Transactions on Medical Imaging* **35**, 1182–1195 (2016).
13. Maier, O., Schröder, C., Forkert, N. D., Martinetz, T. & Handels, H. Classifiers for Ischemic Stroke Lesion Segmentation: A Comparison Study. *PLOS ONE* **10**, e0145118 (2015).
14. Liu, S. *et al.* Multimodal Neuroimaging Feature Learning for Multiclass Diagnosis of Alzheimer’s Disease. *IEEE Trans. Biomed. Eng.* **62**, 1132–1140 (2015).
15. Plis, S. M. *et al.* Deep learning for neuroimaging: a validation study. *Front Neurosci* **8**, (2014).
16. Davatzikos, C. Machine learning in neuroimaging: Progress and challenges. *Neuroimage* **197**, 652–656 (2019).
17. Liu, S. *et al.* Early diagnosis of Alzheimer’s disease with deep learning. in *2014 IEEE 11th International Symposium on Biomedical Imaging (ISBI)* 1015–1018 (IEEE, 2014). doi:10.1109/ISBI.2014.6868045.
18. Suk, H.-I. & Shen, D. Deep learning-based feature representation for AD/MCI classification. in *MICCAI International Conference on Medical Image Computing and Computer-Assisted Intervention* vol. 16 583–590 (2013).
19. Kuang, D., Guo, X., An, X., Zhao, Y. & He, L. Discrimination of ADHD Based on fMRI Data with Deep Belief Network. in *Intelligent Computing in Bioinformatics* (eds. Huang, D.-S., Han, K. & Gromiha, M.) 225–232 (Springer International Publishing, 2014). doi:10.1007/978-3-319-09330-7_27.
20. 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 (2017).
21. Hon, M. & Khan, N. Towards Alzheimer’s Disease Classification through Transfer Learning. *arXiv:1711.11117 [cs]* (2017).
22. Sarraf, S., DeSouza, D. D., Anderson, J. & Tofighi, G. DeepAD: Alzheimer’s Disease Classification via Deep Convolutional Neural Networks using MRI and fMRI. *bioRxiv* 070441 (2017) doi:10.1101/070441.
23. Wu, C. *et al.* Discrimination and conversion prediction of mild cognitive impairment using convolutional neural networks. *Quant Imaging Med Surg* **8**, 992–1003 (2018).
24. Islam, J. & Zhang, Y. Brain MRI analysis for Alzheimer’s disease diagnosis using an ensemble system of deep convolutional neural networks. *Brain Inform* **5**, 2 (2018).
25. Esmailzadeh, S., Yang, Y. & Adeli, E. End-to-End Parkinson Disease Diagnosis using Brain MR-Images by 3D-CNN. *arXiv:1806.05233* (2018).
26. Sivaranjini, S. & Sujatha, C. M. Deep learning based diagnosis of Parkinson’s disease using convolutional neural network. *Multimedia Tools and Applications* (2019) doi:10.1007/s11042-019-7469-8.
27. Wen, J. *et al.* Convolutional neural networks for classification of Alzheimer’s disease: Overview and reproducible evaluation. *Medical Image Analysis* **63**, 101694 (2020).

28. Blum, A., Kalai, A. & Langford, J. Beating the hold-out: bounds for K-fold and progressive cross-validation. in *Proceedings of the twelfth annual conference on Computational learning theory - COLT '99* 203–208 (ACM Press, 1999). doi:10.1145/307400.307439.
29. Yadav, S. & Shukla, S. Analysis of k-Fold Cross-Validation over Hold-Out Validation on Colossal Datasets for Quality Classification. in *2016 IEEE 6th International Conference on Advanced Computing (IACC)* 78–83 (IEEE, 2016). doi:10.1109/IACC.2016.25.
30. Yagis, E., De Herrera, A. G. S. & Citi, L. Generalization Performance of Deep Learning Models in Neurodegenerative Disease Classification. in *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* 1692–1698 (IEEE, 2019). doi:10.1109/BIBM47256.2019.8983088.
31. Marcus, D. S. *et al.* Open Access Series of Imaging Studies (OASIS): cross-sectional MRI data in young, middle aged, nondemented, and demented older adults. *J Cogn Neurosci* **19**, 1498–1507 (2007).
32. Petersen, R. C. *et al.* Alzheimer's Disease Neuroimaging Initiative (ADNI). *Neurology* **74**, 201–209 (2010).
33. Marek, K. *et al.* The Parkinson's progression markers initiative (PPMI) - establishing a PD biomarker cohort. *Ann Clin Transl Neurol* **5**, 1460–1477 (2018).
34. Murad, M. *et al.* Efficient Reconstruction Technique for Multi-Slice CS-MRI Using Novel Interpolation and 2D Sampling Scheme. *IEEE Access* **8**, 117452–117466 (2020).
35. Farooq, A., Anwar, S., Awais, M. & Rehman, S. A deep CNN based multi-class classification of Alzheimer's disease using MRI. in *2017 IEEE International Conference on Imaging Systems and Techniques (IST)* 1–6 (2017). doi:10.1109/IST.2017.8261460.
36. Suk, H.-I., Shen, D., & Alzheimer's Disease Neuroimaging Initiative. Deep Learning in Diagnosis of Brain Disorders. in *Recent Progress in Brain and Cognitive Engineering* (eds. Lee, S.-W., Bülthoff, H. H. & Müller, K.-R.) vol. 5 203–213 (Springer Netherlands, 2015).
37. Kobayashi, S., Kane, T. & Paton, C. The Privacy and Security Implications of Open Data in Healthcare: A Contribution from the IMIA Open Source Working Group. *Yearb Med Inform* **27**, 041–047 (2018).
38. Celi, L. A., Citi, L., Ghassemi, M. & Pollard, T. J. The PLOS ONE collection on machine learning in health and biomedicine: Towards open code and open data. *PLoS ONE* **14**, e0210232 (2019).
39. Varma, S. & Simon, R. Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics* **7**, 91 (2006).
40. Reunanen, J. Overfitting in Making Comparisons Between Variable Selection Methods. *J. Mach. Learn. Res* **3**, 1371–1382 (2003).
41. Morris, J. C. The Clinical Dementia Rating (CDR): current version and scoring rules. *Neurology* **43**, 2412–2414 (1993).
42. Morris, J. C. *et al.* Mild cognitive impairment represents early-stage Alzheimer disease. *Arch. Neurol.* **58**, 397–405 (2001).
43. Hoehn, M. M. & Yahr, M. D. Parkinsonism: onset, progression and mortality. *Neurology* **17**, 427–442 (1967).
44. Tessa, C. *et al.* Central modulation of parasympathetic outflow is impaired in de novo Parkinson's disease patients. *PLoS ONE* **14**, e0210324 (2019).
45. Han, X. *et al.* Brain extraction from normal and pathological images: A joint PCA/Image-Reconstruction approach. *Neuroimage* **176**, 431–445 (2018).
46. Avants, B. B. *et al.* A Reproducible Evaluation of ANTs Similarity Metric Performance in Brain Image Registration. *Neuroimage* **54**, 2033–2044 (2011).

47. Bradski, G. R. & Kaehler, A. *Learning OpenCV: computer vision with the OpenCV library*. (O'Reilly, 2011).
48. Simonyan, K. & Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv:1409.1556 [cs]* (2015).
49. He, K., Zhang, X., Ren, S. & Sun, J. Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. in *2015 IEEE International Conference on Computer Vision (ICCV)* 1026–1034 (IEEE, 2015). doi:10.1109/ICCV.2015.123.
50. Glorot, X. & Bengio, Y. Understanding the difficulty of training deep feedforward neural networks. in *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics* 249–256 (2010).
51. Lin, M., Chen, Q. & Yan, S. Network In Network. *arXiv:1312.4400 [cs]* (2014).
52. Szegedy, C. *et al.* Going deeper with convolutions. in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 1–9 (2015). doi:10.1109/CVPR.2015.7298594.
53. Hastie, T., Tibshirani, R. & Friedman, J. H. *The elements of statistical learning: data mining, inference, and prediction*. (Springer, 2009).
54. Lemm, S., Blankertz, B., Dickhaus, T. & Müller, K.-R. Introduction to machine learning for brain imaging. *NeuroImage* **56**, 387–399 (2011).
55. Kingma, D. P. & Ba, J. Adam: A Method for Stochastic Optimization. *CoRR* (2015).
56. Chollet, F. & others. *Keras*. (2015).
57. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* **12**, 2825–2830 (2011).
58. Gunawardena, K. A. N. N. P., Rajapakse, R. N. & Kodikara, N. D. Applying convolutional neural networks for pre-detection of alzheimer's disease from structural MRI data. in *2017 24th International Conference on Mechatronics and Machine Vision in Practice (M2VIP)* 1–7 (2017). doi:10.1109/M2VIP.2017.8211486.
59. Jain, R., Jain, N., Aggarwal, A. & D, J. Convolutional Neural Network based Alzheimer's Disease Classification from Magnetic Resonance Brain Images. *Cognitive Systems Research* **57**, (2019).
60. Khagi, B., Lee, B., Pyun, J.-Y. & Kwon, G.-R. CNN Models Performance Analysis on MRI images of OASIS dataset for distinction between Healthy and Alzheimer's patient. in *2019 International Conference on Electronics, Information, and Communication (ICEIC)* 1–4 (IEEE, 2019). doi:10.23919/ELINFOCOM.2019.8706339.
61. Wang, S., Shen, Y., Chen, W., Xiao, T.-F. & Hu, J. Automatic Recognition of Mild Cognitive Impairment from MRI Images Using Expedited Convolutional Neural Networks. in *ICANN* (2017). doi:10.1007/978-3-319-68600-4_43.
62. Puranik, M., Shah, H., Shah, K. & Bagul, S. Intelligent Alzheimer's Detector Using Deep Learning. in *2018 Second International Conference on Intelligent Computing and Control Systems (ICICCS)* 318–323 (IEEE, 2018). doi:10.1109/ICCONS.2018.8663065.
63. Basheera, S. & Sai Ram, M. S. Convolution neural network–based Alzheimer's disease classification using hybrid enhanced independent component analysis based segmented gray matter of T2 weighted magnetic resonance imaging with clinical valuation. *Alzheimer's & Dementia: Translational Research & Clinical Interventions* **5**, 974–986 (2019).
64. Basaia, S. *et al.* Automated classification of Alzheimer's disease and mild cognitive impairment using a single MRI and deep neural networks. *NeuroImage: Clinical* **21**, 101645 (2019).
65. Ramzan, F. *et al.* A Deep Learning Approach for Automated Diagnosis and Multi-Class Classification of Alzheimer's Disease Stages Using Resting-State fMRI and Residual Neural Networks. *J Med Syst* **44**, 37 (2019).

66. Raza, M. *et al.* Diagnosis and Monitoring of Alzheimer's Patients Using Classical and Deep Learning Techniques. *Expert Systems With Applications* **136**, 353–364 (2019).
67. Pathak, K. C. & Kundaram, S. S. Accuracy-Based Performance Analysis of Alzheimer's Disease Classification Using Deep Convolution Neural Network. in *Soft Computing: Theories and Applications* (eds. Pant, M., Kumar Sharma, T., Arya, R., Sahana, B. C. & Zolfagharinia, H.) vol. 1154 731–744 (Springer Singapore, 2020).
68. Libero, L. E., DeRamus, T. P., Lahti, A. C., Deshpande, G. & Kana, R. K. Multimodal neuroimaging based classification of autism spectrum disorder using anatomical, neurochemical, and white matter correlates. *Cortex* **66**, 46–59 (2015).
69. Zhou, Y., Yu, F. & Duong, T. Multiparametric MRI Characterization and Prediction in Autism Spectrum Disorder Using Graph Theory and Machine Learning. *PLoS One* **9**, (2014).
70. Lui, Y. W. *et al.* Classification algorithms using multiple MRI features in mild traumatic brain injury. *Neurology* **83**, 1235–1240 (2014).
71. Hasan, A. M., Jalab, H. A., Meziane, F., Kahtan, H. & Al-Ahmad, A. S. Combining Deep and Handcrafted Image Features for MRI Brain Scan Classification. *IEEE Access* **7**, 79959–79967 (2019).
72. Al-Khuzai, F. E. K., Bayat, O. & Duru, A. D. Diagnosis of Alzheimer Disease Using 2D MRI Slices by Convolutional Neural Network. *Appl Bionics Biomech* **2021**, 6690539 (2021).
73. Billones, C. D., Demetria, O. J. L. D., Hostallero, D. E. D. & Naval, P. C. DemNet: A Convolutional Neural Network for the detection of Alzheimer's Disease and Mild Cognitive Impairment. in *2016 IEEE Region 10 Conference (TENCON)* 3724–3727 (IEEE, 2016). doi:10.1109/TENCON.2016.7848755.

Figures

Figure 1

Schematic diagram of the overall T1-weighted MRI data processing and validation scheme. First, a pre-processing stage included co-registration to a standard space, skull-stripping and slices selection based on entropy calculation. Then, CNNs model's training and validation have been performed on each dataset in a nested CV loop using two different data split strategies: a) subject-level split, in which all the slices of a subject have been placed either in the training or in the test set, avoiding any form of data leakage; b) slice-level split, in which all the slices have been pooled together prior to CV, then split randomly into training and test set.

Figure 2

The two different networks based on the VGG16 architecture are shown. Each colored block of layers illustrates a series of convolutions. (a) The first model, named as VGG16-v1 consists of five convolutional blocks followed by three fully connected layers. Only the last three fully connected layers are fine-tuned. (b) On the other hand, the second model, VGG16-v2, has 5 convolutional blocks followed by a global average pooling layer and all the layers are fine-tuned.

Figure 3

A modified ResNet-18 architecture with an average pooling layer at the end is shown. The upper box represents a residual learning block with an identity shortcut. Each layer is denoted as (filter size, # channels); layers labeled as “frozen” indicates that the weights are not updated during backpropagation, whereas when they are labeled as “fine-tuned” they are updated. The identity shortcuts can be directly used when the input and output are of the same dimensions (solid line shortcuts) and when the dimensions increase (dotted line shortcuts). ReLU = rectified linear unit.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [SupplementaryInformation.pdf](#)