A Practical Alzheimer Disease Classifier via Brain Imaging-Based Deep Learning on 85,721 Samples

Bin Lu
University of Chinese Academy of Sciences
Hui-Xian Li
University of Chinese Academy of Sciences
Zhi-Kai Chang
University of Chinese Academy of Sciences
Le Li
Beijing Language and Culture University
Ning-Xuan Chen
University of Chinese Academy of Sciences
Zhi-Chen Zhu
University of Chinese Academy of Sciences
Hui-Xia Zhou
University of Chinese Academy of Sciences
Xue-Ying Li
University of Chinese Academy of Science
Yu-Wei Wang
University of Chinese Academy of Sciences
Shi-Xian Cui
University of Chinese Academy of Science
Zhao-Yu Deng
University of Chinese Academy of Sciences
Zhen Fan
Fudan University
Hong Yang
Zhejiang University
Xiao Chen
University of Chinese Academy of Sciences
Paul M. Thompson
University of Southern California
Francisco Xavier Castellanos
NYU Grossman School of Medicine
Chao-Gan Yan (ycg.yan@gmail.com)
University of Chinese Academy of Sciences

Research Article

Keywords: Alzheimer's disease, convolutional neural network, magnetic resonance brain imaging, sex difference, transfer learning

Posted Date: December 14th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-1156067/v1

Abstract

Background

Beyond detecting brain lesions or tumors, comparatively little success has been attained in identifying brain disorders such as Alzheimer's disease (AD), based on magnetic resonance imaging (MRI). Many machine learning algorithms to detect AD have been trained using limited training data, meaning they often generalize poorly when applied to scans from previously unseen populations. Therefore, we built a practical brain MRI-based AD diagnostic classifier using deep learning/transfer learning on dataset of unprecedented size and diversity.

Methods

A retrospective MRI dataset pooled from more than 217 sites/scanners constituted the largest brain MRI sample to date (85,721 scans from 50,876 participants) between January 2017 and August 2021. Next, a state-of-the-art deep convolutional neural network, Inception-ResNet-V2, was built as a sex classifier with high generalization capability. The sex classifier achieved 94.9% accuracy and served as a base model in transfer learning for the objective diagnosis of AD.

Findings

After transfer learning, the model fine-tuned for AD classification achieved 91.3% accuracy in leave-sites-out cross-validation on the Alzheimer's Disease Neuroimaging Initiative (ADNI, 6,857 samples) dataset and 94.2%/93.6%/90.5% accuracy for direct tests on three unseen independent datasets (AIBL, 669 samples / MIRIAD, 644 samples / OASIS, 1,123 samples). When this AD classifier was tested on brain images from unseen mild cognitive impairment (MCI) patients, MCI patients who finally converted to AD were 3 times more likely to be predicted as AD than MCI patients who did not convert (65.2% vs 20.6%). Predicted scores from the AD classifier showed significant correlations with illness severity.

Interpretation

In sum, the proposed AD classifier could offer a medical-grade marker that have potential to be integrated into AD diagnostic practice.

Introduction

Magnetic resonance imaging (MRI) is widely used in neuroradiology to detect brain lesions including stroke, vascular disease, and tumor tissue. Even so, MRI has been less useful in definitively identifying degenerative diseases including Alzheimer's disease (AD), mainly because signatures of the disease are diffusely found in the images and hard to distinguish from normal aging. Machine learning and deep learning methods have been trained on relatively small datasets, but the limited training data often leads to poor generalization performance on new datasets not used the train the algorithms. In the current study, we aim to create a practical brain imaging-based AD classifier with high generalization capability via learning/transfer learning on a diverse range of large-scale datasets.

In recently updated AD diagnostic criteria, such as those proposed by IWG-2 and NIA-AA, markers such as amyloid measures from cerebrospinal fluid (CSF) and amyloid-sensitive positron emission tomography (PET) have been integrated into the diagnosis AD(1, 2). The diagnostic sensitivity and specificity have been greatly improved by these markers (1, 3). Even so, the invasive nature and somewhat lower availability of these markers limits their application in routine clinical settings. Structural MRI is a more promising candidate for imaging-based auxiliary diagnosis considering its non-invasive nature and wider availability than PET. In addition, well-developed MRI data preprocessing pipelines make it feasible to integrate MRI markers into automatic end-to-end deep learning algorithms. Deep learning has already been successfully deployed in real-world scenarios such as extreme weather condition prediction(4), aftershock pattern prediction(5) and automatic speech recognition(6). In clinical scenarios, convolutional neural networks (CNN) – a widely-used architecture that is well-suited for image-based deep learning has been successfully used for objective diagnosis of retinal diseases(7), skin cancer(8), and for breast cancer screening(9).

However, prior attempts at MRI-based AD diagnosis have yet to attain clinical utility. A major challenge for brain MRI-based algorithms, especially if they are trained on limited data, is their failure to generalize. Brain imaging data varies depending on scanner characteristics such as scanner vendor, magnetic field strength, head coil hardware, pulse sequence, applied gradient fields, reconstruction methods, scanning parameters, voxel size, field of view, etc. Participants also differ in sex, age, race/ethnicity, and education. Robust methods need to work well on diverse populations. These variations in the scans - and in the populations studied – make it hard for a brain imaging-based classifier trained on data from a single site (or a few sites) to generalize to data from unseen sites/scanners. This has prevented brain imaging-based classifiers from becoming practically useful in clinical settings. Most brain MRI-based studies either did not include independent validation(10-12) or did not achieve satisfactory performance in independent validations(13). In fact, reviews of brain imaging-based AD classifiers suggest that most machine learning methods have been trained on hundreds of samples, only 2 out of 81 studies(14), 0 out of 16 studies(15), and 6 out of 114 studies(16) (of those reviewed in recent systematic reviews) included independent dataset validations which led to challenge to generalization ability of these models.

Therefore, one bottleneck in developing a practical brain imaging-based classifier is the variety and comprehensiveness of training datasets. Directly training AD models on the datasets that only contain several hundred samples may result in overfitting with poor generalization to unseen test data(14). The transfer learning framework has been proposed to solve this problem, by training a model on a certain characteristic for which abundant samples are available, and fine-tuning it to another characteristic, or for similar tasks, in smaller samples(17). There is published evidence that pretrained models can outperformed models trained from scratch in classification accuracy and robustness(18, 19). In the medical imaging area, transfer learning has been successfully applied for diagnosing retinal disease(7) and skin cancer(8). Nonetheless, in the brain imaging field, no studies have fully implemented tens of thousands of openly shared brain images to promote the generalizability of an AD classifier. Thus, in the current study, we used the largest and most diverse sample to date (n =

85,721 from more than 217 sites/scanners, see Table 1) to pre-train a brain imaging-based classifier, to ensure high generalizability. We prefer sex classifier rather than age predictor as base model in transfer learning, because the age prediction error may contain biological meaning (e.g. inflated predicted age may indicate accelerated aging (20)). It's hard to measure the true performance of the age predictor but the sex for participants are more stable to be classified. After that, the pre-trained sex classifier was fine-tuned for AD classification and was validated through leave-sites-out cross-validation and three independent validation.

The goal of the present study was to build a practical AD classifier with high generalization capability. We incorporated three design features to improve the clinical utility of the method. First, we trained and tested the algorithm on a dataset of unprecedented size and diversity - from more than 217 sites/scanners - the variety of training samples critical for improving the generalization capability of the models. Secondly, a rigorous leave-datasets/sites-out cross-validation and independent validations were implemented to make sure that the classifier accuracy would be robust to site/scanner variability. Thirdly, compared to 2D modules (feature detectors) typically used in CNNs for natural images, fully 3D convolution filters in the present study were used capture more sophisticated and distributed spatial features for diagnostic classification. We also openly share our preprocessed data, trained model, code, and have built an online predicting website for anyone interested in testing our classifier.

Methods

Data acquisition. We submitted data access applications to nearly all the open-access brain imaging data archives and received permissions from the administrators of 34 datasets. The full dataset list is shown in table 1. Deidentified data were contributed from datasets approved by local Institutional Review Boards. The reanalysis of these data was approved by the Institutional Review Board of Institute of Psychology, Chinese Academy of Sciences. All participants had provided written informed consent at their local institution. All 50,876 participants (contributing 85,721 samples) had at least one session with a T1-weighted structural brain image and information on their sex and age. For participants with multiple sessions of structural images, each image was considered as an independent sample for data augmentation in training. Importantly, scans from the same person were never split into training and testing sets, as that could artifactually inflate performance.

Table 1: **The datasets used in the present study**

Full Name of Dataset	Number of T1 Scans	T1 Scans after QC	Age (mean±std)	Number of Subjects	Number of Sites	Manufacturers	Field Strength
Adolescent Brain Cognition Development	31176	30222	13.76±10.08	11875	21	SIE/PHI/GE	3T
UK Biobank	20124	19744	63.1±7.46	20124	4	SIE	3T
Alzheimer's Disease Neuroimaging Initiative	16596	16431	74.97±7.4	2546	57	SIE/PHI/GE	3T/1.5T
Open Access Series of Imaging Studies	3150	3099	67.54±20.64	1664	(5)	SIE	3T/1.5T
REST-meta-MDD sample	2380	2363	36.2±15.11	2380	17	SIE/PHI/GE	3T/1.5T
Brain Genomics Superstruct Project	1570	1552	21.54±2.89	1570	2	SIE	3T
Human Connectome Project	1267	1220	/	1267	1	SIE	7T/3T
Autism Brain Imaging Data Exchange	1102	1073	17.09±8.06	1102	17	SIE/PHI/GE	3T/1.5T
Autism Brain Imaging Data Exchange II	1043	1019	15.16±9.39	1043	19	SIE/PHI/GE	3T/1.5T
1000 Functional Connectomes Project	897	864	25.8±10.76	897	33	SIE/PHI/GE	3T/1.5T
ADHD-200 Sample	876	864	12.35±3.28	876	8	SIE/PHI	3T/1.5T
Consortium for Reliability and Reproducibility	714	691	23.45±12.31	715	2	SIE/GE	3T
Cambridge Centre for Ageing Neuroscience (Cam-CAN)	652	523	54.36±18.55	652	1	SIE	3T
Enhanced Nathan Kline Institute - Rockland Sample	646	616	38.63±21.21	646	1	SIE	3T
Southwest University Longitudinal Imaging Multimodal	586	581	20.1±1.3	586	1	SIE	3T
Child Mind Institute Healthy Brain Network	572	506	10.74±3.65	572	3	SIE	3T/1.5T
Establishing Moderators and Biosignatures of Antidepressant Response in Clinic Care	540	523	/	540	4	SIE/PHI/GE	3T
Southwest University Adult Lifespan Dataset	493	483	45.16±17.45	493	1	SIE	3T
Max Planck Institute Leipzig Mind-Brain-Body Dataset	316	291	/	316	1	SIE	3T
Beijing Enhanced Sample	180	176	21.22±1.94	180	1	SIE	3T
Nathan Kline Institute - Rockland Sample	167	151	35.59±20.71	167	1	SIE	3T
The Center for Biomedical Research Excellence	147	137	/	147	1	SIE	3T
The Age-ility Project	110	105	21.87±5.39	110	1	SIE	3T
Parkinson's Disease Datasets	68	65	66.18±7.58	68	2	SIE	3T/1.5T
Power et al., 2012 Neuroimage Sample	63	62	14.25±6.05	63	1 (2)	SIE	3T
NYU Institute for Pediatric Neuroscience	47	45	30.4±8.98	47	1	SIE	3T
Beijing Eyes Open Eyes Closed Sample	46	44	22.54±2.18	46	1	SIE	3T
Multi-Modal MRI Reproducibility Resource	42	42	31.76±9.35	42	1	PHI	3T
Adelstein et al., 2011, PLoS ONE Sample	39	36	29.59±8.38	39	1	SIE	3T
Cleveland CCF	31	29	43.55±11.14	31	1	SIE	3T
Virginia Tech Carilion Research Institute	25	24	26.84±8.17	25	1 (3)	SIE	3T
Beijing Short TR Sample	24	23	23.71±6.74	24	1	SIE	3T
FIND Lab sample	13	12	24.08±3.73	13	1	GE	3T
The Midnight Scan Club dataset	10	10	29.1±3.35	10	1	SIE	3T
	85712	83735		50876	217		

Note: Abbreviation: SIE = Siemens, Phi = Philips, GE = General Electric. Numbers in the bracket indicate the number of scanners.

MRI preprocessing. We did not feed raw data into the classifier for training, but used accepted pre-processing pipelines that are known to generate valuable features from the brain scans. The brain structural data were segmented and normalized to acquire grey matter density (GMD) and grey matter volume (GMV) maps. Specifically, we used the voxel-based morphometry (VBM) analysis module within Data Processing Assistant for Resting-State fMRI (DPARSF)(21), which is based on SPM(22), to segment individual T1-weighted images into grey matter, white matter, and cerebrospinal fluid (CSF). Then, the segmented images were transformed from individual native space to MNI space (a coordinate system created by Montreal Neurological Institute) using the Diffeomorphic

Anatomical Registration Through Exponentiated Lie algebra (DARTEL) tool(23). Two voxel-based structural metrics, GMD and GMV were fed into the deep learning classifier as two features for each participant. GMD is the output of unmodulated tissue segmentation map in the MNI space. GMV is calculated by multiplying the voxel value in GMD by the Jacobian determinants derived from the spatial normalization step (modulated) (24).

Quality control. Poor quality raw structural images would produce distorted GMD and GMV maps during segmentation and normalization. To remove such participants from affecting the training of the classifiers, we excluded participants in each dataset with a spatial correlation exceeding the threshold defined by (mean - 2SD) of the Pearson's correlation between each participant's GMV map and the grand mean GMV template. The grand mean GMV template was generated by randomly selecting 10 participants from each dataset and averaging the GMV maps of all these 340 (from 34 datasets) participants. All these participants were visually checked for image quality. After quality control, 83,735 samples were retained for classifier training (figure S1).

Deep learning: classifier training and testing for sex. We trained a 3-dimensional Inception-ResNet-v2(25) model adopted from its 2-dimensional version in the Keras built-in application (see figure 1A for its structure). This is a state-of-the-art model in pattern recognition, and it integrates two classical series of CNN models, Inception and ResNet. We replaced the convolution, pooling, and normalization modules with their 3-dimensional versions and adjusted the number of layers and convolutional kernels to make them suitable for 3-dimensional MRI inputs (e.g., GMD and GMV as different input channels). The present model consists of one stem module, three groups of convolutional modules (Inception-ResNet-A/B/C) and two reduction modules (Reduction-A/B). The model can take advantage of convolutional kernels with different shapes and sizes, and can extract features of different sizes. The model also can mitigate vanishing gradients and exploding gradients by adding residual modules. We utilized the Keras built-in stochastic gradient descent optimizer with learning rate = 0.01, Nesterov momentum = 0.9, decay = 0.003 (e.g., learn rate = learn rate_0 × (1 / (1 + decay × batch))). The loss function was set to binary cross-entropy. The batch size was set to 24 and the training procedure lasted 10 epochs for each fold. To avoid potential overfitting, we randomly split 600 samples out of the training sample as a validation sample and set a checking point at the end of every epoch. We saved the model in which the epoch classifier showed the lowest validation loss. Thereafter, the testing sample was fed into this model to test the classifier.

While training the sex classifier, random cross-validation may share participants from the same sites between training and testing samples, so the model may not generalize well to datasets from unseen sites due to the site information leakage during training. To ensure generalizability, we used cross-dataset validation. In the testing phase, all the data from a given dataset would never be seen during the classifier training phase. This also ensured the data from a given site (and thus a given scanner) were unseen by the classifier during training (see figure1B for an illustration). This strict setting can limit classifier performance, but it makes it feasible to generalize to any participant at any site (scanner). Five-fold cross-dataset validation was used to assess classifier accuracy. Of note, 3 datasets were always kept in the training sample due to the massive number of samples: Adolescent Brain Cognition Development (ABCD) (n = 31,176), UK Biobank (n = 20,124), and the Alzheimer's Disease Neuroimaging Initiative (ADNI) (n = 16,596). The remaining 31 datasets were randomly allocated to the training and testing samples. The allocating schemas were the solution that balanced the sample size of 5 folds the best from 10,000 random allocating procedures. Both healthy normal control and brain-related disorder patient samples in 34 datasets were used in training the sex classifier.

Transfer learning: classifier training and testing for AD. After obtaining a highly robust and accurate brain imaging-based sex classifier as a base model, we used transfer learning to further fine-tune the AD classifier. Rather than retaining the intact sophisticated structure of the base model (Inception-ResNet-V2), we only leveraged the pre-trained weights in the stem module and simplified the upper layers (e.g., replacing Inception-ResNet modules with ordinary convolutional layers). The retained bottom structure of the model works as a feature extractor and can take advantage of the massive training of sex classifier. And the pruned upper structure of the AD model can avoid potential overfitting and promote generalizability by reducing the number of parameters (10 million parameters for the AD classifier vs. 54 million parameters for the sex classifier). This derived AD classifier was fine-tuned on the ADNI dataset (2,186 samples from 380 AD patients and 4,671 samples from 698 normal controls (NCs), 76 ± 7 years, 3,493 samples from women). ADNI was launched in 2003 (Principal Investigator: Michael W. Weiner, MD) to investigate biological markers of the progression of MCI and early AD (see www.adni-info.org). We used the Keras built-in stochastic gradient descent optimizer with learning rate = 0.0003, Nesterov momentum = 0.9, decay = 0.002. The loss function was set to binary cross-entropy. The batch size was set to 24 and the training procedure lasted 10 epochs for each fold. Similar to the cross-dataset validation for the sex classifier training, five-fold cross-site validation was used to assess classifier accuracy (see figure 1C for an illustration). By ensuring that the data from a given site (and thus a given scanner) were unseen by the classifier during training, this strict strategy made the classifier generalizable with non-inflated accuracy, thus better simulating realistic clinical applications than traditional five-fold cross-validation. Other than using GMD+GMV as the input in transfer learning, we also used GMD, GMV or z-standardized normalized raw T1-weighted images as the input for the sex/AD classifiers to verify the influence of the input format (table 2). We also trained an age prediction model instead of the sex classifier in transfer learning to verify the influence of the base-model. We used the same structure as the sex classifier, except for adding a fully-connected layer with 128 neurons with "ELU" activation function before the final layer; we also changed the dropout rate from 0.5 to 0.2 following the parameters in the brain age prediction model reported by B. A. Jonsson et al. (20).

Furthermore, to test the generalizability of the AD classifier, we directly tested it on three unseen independent AD samples, i.e., the Australian Imaging, Marker and Lifestyle Flagship Study of Ageing (AIBL)(26), the Minimal Interval Resonance Imaging in Alzheimer's Disease cohort (MIRIAD)(27), and the Open Access Series of Imaging Studies (OASIS)(28). We used an ensemble method, based on averaging the output of 5 AD classifiers in the five-fold cross-validation trained on ADNI, to obtain the final classification for each sample. We used diagnoses provided by the qualified physicians for the AIBL and MIRIAD datasets as the labels of samples (115 samples from 82 AD patients and 554 samples from 324 NCs in AIBL, 74 ± 7 years, 374 samples from women; 409 samples from 46 AD patients and 235 samples from 23 NCs in MIRIAD, 70 ± 7 years, 358 samples from women). As OASIS did not specify the criteria for an AD diagnosis, we adopted criteria of MMSE and clinical dementia rating (CDR) modified from ADNI-1 protocol manual to define AD and NC samples. To be specific, criteria for AD are (1) MMSE <= 22 and (2) CDR >= 1.0, and criteria for NC are (1) MMSE > 26 and (2) CDR = 0. Thus, we tested the model on 137 samples from 34 AD patients and 986 samples from 213 NC participants in the OASIS dataset after quality control, 75 ± 10 years, 772 samples from women. Of note, the scanning conditions and recruitment criteria of these independent datasets differed much more than variations among different ADNI sites (where scanning and recruitment was deliberately coordinated), so we expected the AD classifier to achieve lower performance.

We further investigated whether the AD classifier could predict disease progression in people with mild cognitive impairment (MCI). MCI is a syndrome defined as relative cognitive decline without symptoms interfering with daily life; even so, more than half of MCI patients progress to dementia within 5 years(29). The stable MCI (sMCI) samples were defined as “scans from a individuals who was once diagnosed as MCI in any phase of ADNI and has not progressed to AD by the end of the ADNI follow-up”, and the progressive MCI (pMCI) samples were defined as “scans from a participant who was once diagnosed as MCI in any phase of ADNI and who has progressed to AD”. The scans labeled as “conversion” or “AD” (after conversion) for pMCI and the last scan for sMCI were excluded in the present study for precision. We screened imaging records of the MCI patients who converted to AD later in the ADNI 1/2/’GO’ phases, and collected 2,371 images from 243 participants labeled as ‘pMCI’. We also assembled 4,018 samples from 524 participants labeled ‘sMCI’ without later progression for contrast. We directly fed all these MCI images into the AD classifier without further fine-tuning, thus evaluating the performance of the AD classifier on unseen MCI information.

Interpretation of the deep learning classifiers.

To better understand the brain imaging-based deep learning classifier, we calculated occlusion maps for the classifiers. We repeatedly tested the images in testing sample using the model with the highest accuracy within the 5 folds, while successively masking brain areas (volume = 18mm*18mm*18mm, step = 9mm) of all input images. The accuracy achieved on “intact” samples by the classifier minus accuracy achieved on “defective” samples indicated the “importance” of the occluded brain area for the classifier. The occlusion maps were calculated for both sex and AD classifiers. To investigate the clinical significance of the output of the AD classifier, we calculated the Spearman’s correlation coefficient between the predicted scores and MMSE scores of AD, NC, and MCI samples. We also used general linear models (GLM) to verify whether the predicted scores (or MMSE score) showed a group difference between people with sMCI and pMCI. The age and sex information of MCI participants was included in this GLM as covariates. We selected the T1-weighted images from the first visit for each MCI subject and finally collected data from 243 pMCI patients and 524 sMCI patients.

Results

3.1. Large-Scale Brain imaging Data

Only brain imaging data with enough size and variety can make deep learning accurate and robust enough to build a practical classifier. We received permissions from the administrators of 34 datasets (85,721 samples of 50,876 participants from more than 217 sites/scanners, see Table 1; there were no application requirements for some datasets). After quality control, all these samples were used to pre-train the stem module to achieve better generalization for further AD classifier training. The T1-weighted images were collected through Magnetization-Prepared Rapid Gradient-Echo Imaging (MPRAGE) or Inversion Recovery Fast Spoiled Gradient Recalled Echo (IR-FSPGR) sequences of 1.5 tesla or 3 tesla MR scanners. The raw acquisition voxel sizes ranged from 0.7mm×0.7mm×0.7mm to 1.3mm×1.3mm×1.2mm. For the further fine-tuning of the AD classifier, ADNI, AIBL, MIRIAD, and OASIS were selected to train and test the model.

3.2. Performance of the sex classifier

We trained a 3-dimensional Inception-ResNet-v2 model adapted from its 2-dimensional version in the Keras built-in application (see figure 1A for structure). As noted in the Methods, we did not feed raw data into the classifier for training, but used prior knowledge regarding helpful analytic pipelines. Grey matter density (GMD) and grey matter volume (GMV) maps were treated as different input channels for models. To ensure generalizability, five-fold cross-dataset validation was used to assess classifier accuracy. The five-fold cross-dataset validation accuracies were: 94.8%, 94.0%, 94.8%, 95.7%, and 95.8%. Taken together, accuracy was 94.9% in testing samples when pooling results across the five folds. The area under the curve (AUC) of the receiver operating characteristic (ROC) curve reached 0.981 (**figure 2**). In short, our model can classify the sex of a participant based on brain structural imaging data from anyone and any scanner with an accuracy of about 95%. Interested readers can test this model on our online prediction website (http://brainimagenet.org).

3.3. Performance of the AD classifier

After creating a practical brain imaging-based classifier for sex with high cross-dataset accuracy, we used transfer learning to see if we could classify patients with AD. The AD classifier achieved an average accuracy of 91.3% (accuracy = 93.2%, 90.3%, 92.0%, 94.4%, and 86.7% in 5 cross-site folds) in the test samples. Average sensitivity and specificity were 0.848 and 0.943, respectively. The ROC AUC reached 0.962 when results from the 5 testing samples were taken together (see figure 3 and table 2). The AD classifier achieved an average accuracy of 91.4% in 3T field strength MR testing samples and achieved an average accuracy of 91.1% in 1.5T MR testing samples. The accuracy in 3T MR testing sample was not significantly different from that of 1.5T MR testing sample (p = 0.316, statistical examined by permutation test of randomly allocating the testing samples into 1.5T group or 3T group and calculated the accuracy difference between the two groups for 100,000 times, figure S2).

To test the generalizability of the AD classifier, we applied it to unseen independent AD datasets, i.e., AIBL, MIRIAD, and OASIS. The AD classifier achieved 94.2% accuracy in AIBL with 0.97 AUC (figure 3D). Sensitivity and specificity were 0.881 and 0.954, respectively. The AD classifier achieved 93.6% accuracy in MIRIAD with 0.993 AUC (figure 3E). Sensitivity and specificity were 0.897 and 1.000, respectively. The AD classifier achieved 90.5% accuracy in OASIS with 0.969 AUC (figure 3D). Sensitivity and specificity were 0.903 and 0.905, respectively.

Table 2: **Performance of the Alzheimer's disease classifier**

	Accuracy				AUC				Sensitivity				Specificity		
Input Feature	ADNI	AIBL	MIRIAD	OASIS	ADNI	AIBL	MIRIAD	OASIS	ADNI	AIBL	MIRIAD	OASIS	ADNI	AIBL	MIRIAD
Sex classifier as base model															
GMV+GMD	0.913	0.942	0.936	0.905	0.962	0.970	0.993	0.969	0.848	0.881	0.897	0.903	0.943	0.954	1.000
GMD	0.879	0.941	0.944	0.906	0.940	0.952	0.990	0.945	0.775	0.802	0.959	0.791	0.932	0.967	0.919
GMV	0.903	0.944	0.917	0.861	0.960	0.965	0.991	0.949	0.830	0.861	0.895	0.866	0.936	0.960	0.953
T1-weighted	0.886	0.947	0.938	0.730	0.928	0.961	0.991	0.940	0.766	0.822	0.936	0.918	0.910	0.971	0.940
Age predictor as base model															
GMV+GMD	0.901	0.950	0.942	0.892	0.957	0.965	0.995	0.965	0.814	0.881	0.907	0.866	0.942	0.964	1.000

The ADNI dataset contained 2,186 AD samples and 4,671 NC samples. The AIBL dataset contained 115 AD samples and 554 NC samples. The MIRIAD dataset contained 409 AD samples and 235 NC samples. The OASIS dataset contained 137 AD samples and 986 NC samples. The sample sizes shown here are the numbers of T1-weighted brain MRI scans. Abbreviations: AD = Alzheimer's disease participants, NC = normal control participants, GMD = grey matter density map, GMV = grey matter volume map, T1-weighted = normalized 3D T1-weighted structural image, AUC = area under the curve.

Importantly, although the AD classifier is agnostic to brain imaging data of MCI, we directly tested it on the MCI dataset in ADNI to see if it has the potential to predict the progression of MCI to AD. The idea behind this test is that even though people with MCI do not yet have AD, their scans may appear closer to the AD class learned by the deep learning model. In the end, 65.2% of pMCI patients were predicted as closer to the AD class but only 20.4% of sMCI patients were predicted as having AD by the AD classifier (figure 3F). If the percentage of pMCI patients who were predicted as AD was considered as sensitivity and the percentage of sMCI patients who were predicted as AD was considered as 1-specificity, the AUC of ROC curve for AD classifier reached 0.82. These results suggest that the classifier is practical for screening MCI patients who have a higher risk of progression to AD. In sum, we believe our AD classifier can provide important insights relevant to computer-aided diagnosis and prediction of AD, and we have freely provided it on the website http://brainimagenet.org. Importantly, classification results by the online classifier should be interpreted with caution, as they cannot replace diagnosis by licensed clinicians.

As a supplementary analysis, we also trained the AD classifier that kept the intact structure of the base model in transfer learning (figure S3). The performance of the proposed model was comprehensively inferior to the optimized AD classifier. The “intact” AD classifier achieved an average accuracy of 88.4% with 0.938 AUC in the ADNI test samples (figure S4). Average sensitivity and specificity were 0.814 and 0.917, respectively. When tested on independent samples, the AD classifier achieved 91.2% accuracy in AIBL with 0.948 AUC. Sensitivity and specificity were 0.851 and 0.924, respectively. The AD classifier achieved 93.9% accuracy in MIRIAD with 0.995 AUC. Sensitivity and specificity were 0.905 and 0.996, respectively (figure S5). The AD classifier achieved 86.1% accuracy in OASIS with 0.921 AUC. Sensitivity and specificity were 0.789 and 0.881, respectively. When tested on MCI samples, 63.2% of pMCI patients were predicted as having AD and only 22.1% of sMCI patients was predicted as having AD by the AD classifier.

3.4. Interpretation of the deep learning classifiers

To better understand the brain imaging-based deep learning classifier, we calculated occlusion maps for the classifiers. The occlusion map showed that hypothalamus, superior vermis, pituitary, thalamus, amygdala, putamen, accumbens, hippocampus, and parahippocampal gyrus played critical roles in predicting sex (figure 4A). The occlusion map for the AD classifier highlighted that the hippocampus and parahippocampal gyrus - especially in the left hemisphere - played unique roles in predicting AD (figure 4B, figure S6).

To investigate the clinical significance of the output of the AD classifier, we calculated the Spearman’s correlation coefficient between the predicted scores by the classifier and MMSE scores in AD, NC, and MCI samples, although the classifier had not been trained for MMSE scores before. This analysis confirmed significant negative correlations between the predicted scores and MMSE scores for AD (r = -0.37, $p < 1 \times 10^{-55}$), NC (r = -0.11, $p < 1 \times 10^{-11}$), MCI (r = -0.52, $p < 1 \times 10^{-307}$), and the overall samples (r = -0.64, $p < 1 \times 10^{-307}$) (figure 5, figure S7). As lower MMSE scores indicate more severe cognitive impairment for AD and MCI patients, we confirmed that the more severe the disease, the higher the predicted score by the classifier. In addition, both the predicted scores and MMSE scores showed significant differences between pMCI and sMCI (predicted scores: t = 13.88, $p < 0.001$, Cohen’s d = 1.08; MMSE scores: t = -9.42, $p < 0.01$, Cohen’s d = -0.73, figure S8). Importantly, the effect size of the predicted scores by the classifier is much larger than the behavioral measure (MMSE scores).

Discussion

Using an unprecedentedly diverse brain imaging sample, we pre-trained a sex classifier with about 95% accuracy which served as a base-model for transfer learning to promote model generalization capability. After transfer learning, the model fine-tuned to AD achieved 91.3% accuracy in stringent leave-sites-out cross-validation and 94.2/93.6%/90.5% accuracy for direct tests on unseen independent datasets. Predicted scores from the AD classifier showed significant negative correlations with the severity of illness. The AD classifier also showed the potential to predict the prognosis of MCI patients.

The high accuracy and generalizability of our deep neural network classifiers demonstrate that brain imaging dose have practical potential for auxiliary diagnosis. One of the most prominent advantages of the present protocol is its outstanding generalizability, as validated by leave-sites-out validations and three independent-dataset validations. Performance of the AD classifier remained consistent despite considerable scanner/participant variations across datasets. For example, the accuracies always exceeded 90% and AUCs always exceeded 0.96 in all four datasets. In addition, all the specificities (all exceeded 0.9) were slightly higher than the sensitivities (all exceeded 0.84) in four datasets. The present model outperformed the models in the latest studies whose accuracy range from 72.3% to 95%(30) or from 77% to 87%(13) using the same independent datasets (e.g. AIBL, MIRIAD, and OASIS). It is critical to keep the sensitivity-specificity tradeoff consistent by default, so that the physicians at a different hospital can have consistent expectations for the classification tendencies of the classifier. In addition, the analogous accuracies achieved on 1.5T and 3T MR ADNI imaging data further supported the robustness of the present classifier. Of note, the output of the deep neural network model is a continuous variable, so the threshold can be adjusted to change the sensitivity and specificity for certain purpose. For example, when tested on the AIBL dataset, sensitivity and specificity results were 0.881 and 0.954, respectively, as the default threshold was set at 0.5. However, for screening, the false-negative rate should be minimized even at the cost of higher false-positive rates. If we lower the threshold (e.g., to 0.2), sensitivity can be improved to 0.921 at a cost of decreasing specificity to 0.885. Thus, in our freely available AD prediction website, users can obtain continuous outputs and adjust the threshold by themselves.

Except for the feasibility of being integrated into diagnostic criteria, the present AD model also showed potential to predict the progression of MRI patients. First, the present model was able to quantify key disease milestones by predicting disease progression in MCI patients. In fact, people with pMCI were 3 times more likely to be classified as AD than sMCI (65.2% vs 20.4%). Recently, a critical review about predicting the progression of MCI noted that about 40% of studies had methodological issues, such as lack of a test dataset, data-leakage in feature selection or parameter tuning, and leave-one-out validation performance bias(31). The present AD classifier was only trained on AD/NC samples and was not fine-tuned using MCI data, so data leakage was avoided. The estimated true AUC of current published state-of-art classifiers for predicting progression of MCI is about 0.75(14, 31). The proposed AD classifier here outperformed the benchmark considerably (AUC = 0.82). Considering the discouraging clinical trial failures for AD treatments, early identification of people with MCI with potential to progress would help in evaluation of early treatments(32).

Although deep-learning algorithms have often been described as “black boxes” for their poor interpretability, our subsequent analyses showed that the current MRI-based AD marker was in line with former pathological findings and clinical practices. For example, AD induced brain structural changes have been frequently reported by MRI studies. Among all the structural findings, hippocampal atrophy is the most prominent change and is used in imaging assisted diagnosis(33, 34). Neurobiological changes in the hippocampus typically precede progressive neocortical damage and AD symptoms. The convergence of our deep learning system and human physicians on alterations in hippocampal structure for classifying AD patients is in line with the crucial role of the hippocampus in AD. On the other hand, the maximum absolute value was only about 3.5%, which means that even if the most important brain area was eliminated from input, the accuracy only dropped from 91.3% to about 88.8%. Furthermore, brain atrophy in AD has been frequently reported as left lateralized(35, 36). Compared to the un-optimized AD classifier, a slight left hemisphere preference for input features may help explain the improved performance of the optimized AD classifier.

Rather than indiscriminately imitate the structure of the base model in transfer learning, the present AD classifier significantly simplified the model before the fine-tuning procedure. In fact, the performance of the unoptimized AD classifier was far poorer than that of the optimized AD classifier in accuracy, sensitivity, specificity, and in independent validation performance. There is some reported evidence that truncating or pruning models before transfer learning may facilitate the performance of the transferred models(37, 38). As the sample for training the AD classifier is considerably smaller than that used to train the sex classifier, the simplified model structure may help to avoid overfitting and improve generalizability.

By precisely predicting the sex of people, the present study provided evidences for sex differences in human brain. Daphna and colleagues extracted hundreds of VBM features from structural MRI and concluded that “the so-called male/female brain” does not exist as no individual structural feature supports a sexually dimorphic view of human brains(39), which was supported by a recent large-scale review(40). However, human brains may embody sexually dimorphic features in a multivariate manner. The high accuracy and generalizability of the present sex classifier demonstrate that sex is separable in a 1,981,440-dimension (96*120*86*2) feature space. Among those 1,981,440 features, hypothalamus and pituitary played the most critical role in predicting sex. The hypothalamus regulates testosterone secretion through the hypothalamic-pituitary-gonadal axis and thus plays a critical role in brain masculinization(41). Men have significantly larger hypothalamus than women relative to cerebrum size(42). In addition, cerebellum – especially the vermis - showed a remarkable influence in sex classification, which is in line with MRI morphology studies of sex differences in cerebellum(43, 44). Taken together, our machine learning evidence shows that the “male/female brain” does exist, in the sense that accurate classification is possible.

In the deep learning field, the appearance of ImageNet tremendously accelerated the evolution of computer vision(45). It provided large amounts of well-labeled image data for researchers to pre-train their models. Studies have shown that pre-trained models can facilitate the performance and robustness of subsequently fine-tuned models(18). The present study confirms that the “pre-train + fine-tuning” paradigm does work for MRI-based auxiliary diagnosis. Unfortunately, no such well-preprocessed dataset exists in brain imaging domain. As data organization and preprocessing of MRI data require tremendous time, manpower and computational load, these constraints impede scientists from other fields entering brain imaging. Open access to large amounts of preprocessed brain imaging data is fundamental to facilitate the participation of a broader range of researchers. Beyond building and sharing a practical brain imaging-based deep learning classifier, we would openly share all sharable preprocessed data to invite researchers (especially computer scientists) to join the efforts to create predictive models using brain images (Link_To_Be_Added upon publication, preprocessed data of some datasets could not be shared as the

raw data owners do not allow sharing of data derivatives). We anticipate that this dataset may boost the clinical utility of brain imaging as ImageNet has done in computer vision research. We openly share our models to allow other researchers to deploy them (https://github.com/Chaogan-Yan/BrainImageNet). Our code is also openly shared as well, allowing other researchers to replicate the present results and further develop brain imaging-based classifiers based on our existing work. Finally, we have also built a demonstration website for classifying sex and AD (http://brainimagenet.org). Users can upload their own raw T1-weighted or preprocessed GMD and GMV data to make predictions of sex or AD labels in real-time.

Some limitations of the current study should be acknowledged. Considering the lower reproducibility of functional MRI compared to structural MRI, only structural MRI derived images were used in the present deep learning model. Even so, functional measures of physiology and activation may further improve the performance of sex and brain disorder classifiers. In future studies, functional MRI, especially resting-state functional MRI, may provide additional information for model training. Furthermore, with advances in software such as FreeSurfer(46), fmriprep(47), and DPABISurf, surface-based algorithms have shown their superiority when compared with traditional volume-based algorithms(48). Surface-based algorithms are more time consuming to run in terms of computation load, but can provide more precise brain registration and reproducibility. Future studies should take surface-based images as inputs for deep learning models. In addition, the present AD classification model was built based on labels provided by ADNI database. Further study may also benefit by using post-mortem neuropathological data as a gold standard for AD to further advance the clinical value of MRI-based markers.

In summary, we pooled MRI data from more than 217 sites/scanners to constitute the largest brain MRI sample to date, and applied a state-of-the-art architecture deep convolutional neural network, Inception-ResNet-V2, to pre-train a brain image-based classifier. The AD classifier obtained via transfer learning reached high accuracy and sufficient generalizability to be of practical use, demonstrating the feasibility of transfer learning in brain disorder applications. Future work is needed to deploy such a framework for assessment of psychiatric disorders, to predict treatment response, and other aspects of individual differences.

Declarations

Ethics approval and consent to participate

Deidentified data were contributed from datasets approved by local Institutional Review Boards. The reanalysis of these data was approved by the Institutional Review Board of Institute of Psychology, Chinese Academy of Sciences. All participants had provided written informed consent at their local institution.

Consent for publication

Not applicable.

Availability of data and materials

The imaging, phenotype and clinical data used for the training, validation and test sets were applied from the administrators of 34 datasets. The preprocessed brain imaging data will be available on (Link_To_Be_Added upon publication, preprocessed data of some datasets could not be shared as the raw data owners do not allow sharing data derivatives). The code for training and testing the model are openly shared at https://github.com/Chaogan-Yan/BrainImageNet. Demonstration website for classifying sex and AD was available at http://brainimagenet.org.

Competing interests

The authors declare that they have no competing interests.

Funding

This work was supported by the National Key R&D Program of China (grant number: 2017YFC1309902), the National Natural Science Foundation of China (grant number: 81671774, 81630031), the 13th Five-year Informatization Plan of Chinese Academy of Sciences (grant number: XXH13505), the Key Research Program of the Chinese Academy of Sciences (grant NO. ZDBS-SSW-JSC006), Beijing Nova Program of Science and Technology (grant number: Z191100001119104).

Authors' contributions

C.-G.Y. designed the overall experiment. B.L., H.-X.L., L.L., C.-N.X., Z.-H.C., H.-X.L., Z.F., H.Y. and X.C. applied and preprocessed imaging data. H.-X.L. and B.L sorted the phenotype information of datasets. B.L. designed the model architectures and trained the models, Z.-K.C, B.L. and C.-G.Y. built the online classifiers. C.-G.Y. provided technical supports and supervised the project. B.L. and C.-G.Y. wrote the paper, P.M.T edited the paper and suggested supplementary analysis, F.X.C. edited and polished the paper. All authors approved the manuscript and had final responsibility for the decision to submit for publication.

Acknowledgements

Data used in the preparation of this article for training and testing the sex classifier was obtained from the Adolescent Brain Cognitive Development (ABCD) Study (https://abcdstudy.org), held in the NIMH Data Archive (NDA). This is a multisite, longitudinal study designed to recruit more than 10,000 children age s9-10 and follow them over 10 years into early adulthood. The ABCD Study is supported by the National Institutes of Health and additional federal partners under award numbers U01DA041048, U01DA050989, U01DA051016, U01DA041022, U01DA051018, U01DA051037, U01DA050987, U01DA041174, U01DA041106, U01DA041117, U01DA041028, U01DA041134, U01DA050988, U01DA051039, U01DA041156, U01DA041025, U01DA041120, U01DA051038,

U01DA041148, U01DA041093, U01DA041089. A full list of supporters is available at https://abcdstudy.org/federal-partners.html. A listing of participating sites and a complete listing of the study investigators can be found at https://abcdstudy.org/scientists/workgroups/. ABCD consortium investigators designed and implemented the study and/or provided data but did not necessarily participate in analysis or writing of this report. This manuscript reflects the views of the authors and may not reflect the opinions or views of the NIH or ABCD consortium investigators. This research has been conducted using the UK Biobank Resource. Data collection and sharing for the training and testing the sex and AD classifier were 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 used in the preparation of this article were obtained from the MIRIAD database. The MIRIAD investigators did not participate in analysis or writing of this report. The MIRIAD dataset is made available through the support of the UK Alzheimer's Society (Grant RF116). The original data collection was funded through an unrestricted educational grant from GlaxoSmithKline (Grant 6GKC).

References

1. Dubois B, Feldman HH, Jacova C, Hampel H, Molinuevo JL, Blennow K, et al. Advancing research diagnostic criteria for Alzheimer's disease: the IWG-2 criteria. The Lancet Neurology. 2014;13(6):614-29.
2. Jack Jr CR, Albert MS, Knopman DS, McKhann GM, Sperling RA, Carrillo MC, et al. Introduction to the recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. Alzheimer's & dementia. 2011;7(3):257-62.
3. Rice L, Bisdas S. The diagnostic value of FDG and amyloid PET in Alzheimer's disease-A systematic review. Eur J Radiol. 2017;94:16-24.
4. Ham Y-G, Kim J-H, Luo J-J. Deep learning for multi-year ENSO forecasts. Nature. 2019;573(7775):568-72.
5. DeVries PMR, Viegas F, Wattenberg M, Meade BJ. Deep learning of aftershock patterns following large earthquakes. Nature. 2018;560(7720):632-4.
6. Liu W, Wang Z, Liu X, Zeng N, Liu Y, Alsaadi FE. A survey of deep neural network architectures and their applications. Neurocomputing. 2017;234:11-26.
7. Kermany DS, Goldbaum M, Cai W, Valentim CC, Liang H, Baxter SL, et al. Identifying medical diagnoses and treatable diseases by image-based deep learning. Cell. 2018;172(5):1122-31.
8. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature. 2017;542(7639):115-8.
9. McKinney SM, Sieniek M, Godbole V, Godwin J, Antropova N, Ashrafian H, et al. International evaluation of an AI system for breast cancer screening. Nature. 2020;577(7788):89-94.
10. Suk HI, Lee SW, Shen D, Alzheimer's Disease Neuroimaging I. Hierarchical feature representation and multimodal fusion with deep learning for AD/MCI diagnosis. NeuroImage. 2014;101:569-82.
11. Bashyam VM, Erus G, Doshi J, Habes M, Nasralah I, Truelove-Hill M, et al. MRI signatures of brain age and disease over the lifespan based on a deep brain network and 14 468 individuals worldwide. Brain. 2020;143(7):2312-24.
12. Moradi E, Pepe A, Gaser C, Huttunen H, Tohka J, Alzheimer's Disease Neuroimaging I. Machine learning framework for early MRI-based Alzheimer's conversion prediction in MCI subjects. NeuroImage. 2015;104:398-412.
13. Qiu S, Joshi PS, Miller MI, Xue C, Zhou X, Karjadi C, et al. Development and validation of an interpretable deep learning framework for Alzheimer's disease classification. Brain. 2020;143(6):1920-33.
14. Rathore S, Habes M, Iftikhar MA, Shacklett A, Davatzikos C. A review on neuroimaging-based classification studies and associated feature extraction methods for Alzheimer's disease and its prodromal stages. NeuroImage. 2017;155:530-48.
15. Jo T, Nho K, Saykin AJ. Deep Learning in Alzheimer's Disease: Diagnostic Classification and Prognostic Prediction Using Neuroimaging Data. Front Aging Neurosci. 2019;11:220.
16. Ebrahimighahnavieh MA, Luo S, Chiong R. Deep learning to detect Alzheimer's disease from neuroimaging: A systematic literature review. Comput Methods Programs Biomed. 2020;187:105242.
17. Yosinski J, Clune J, Bengio Y, Lipson H, editors. How transferable are features in deep neural networks? Adv Neural Inf Process Syst; 2014.
18. Hendrycks D, Lee K, Mazeika M. Using pre-training can improve model robustness and uncertainty. arXiv preprint arXiv:190109960. 2019.
19. Tajbakhsh N, Shin JY, Gurudu SR, Hurst RT, Kendall CB, Gotway MB, et al. Convolutional neural networks for medical image analysis: Full training or fine tuning? IEEE Trans Med Imaging. 2016;35(5):1299-312.
20. Jonsson BA, Bjornsdottir G, Thorgeirsson TE, Ellingsen LM, Walters GB, Gudbjartsson DF, et al. Brain age prediction using deep learning uncovers associated sequence variants. Nat Commun. 2019;10(1):5409.
21. Yan CG, Zang YF. DPARSF: A MATLAB Toolbox for "Pipeline" Data Analysis of Resting-State fMRI. Front Syst Neurosci. 2010;4:13.

22. Friston KJ, Holmes AP, Worsley KJ, Poline JP, Frith CD, Frackowiak RS. Statistical parametric maps in functional imaging: a general linear approach. Hum Brain Mapp. 1994;2(4):189-210.
23. Goto M, Abe O, Aoki S, Hayashi N, Miyati T, Takao H, et al. Diffeomorphic Anatomical Registration Through Exponentiated Lie Algebra provides reduced effect of scanner for cortex volumetry with atlas-based method in healthy subjects. Neuroradiology. 2013;55(7):869-75.
24. Good CD, Johnsrude IS, Ashburner J, Henson RN, Friston KJ, Frackowiak RS. A voxel-based morphometric study of ageing in 465 normal adult human brains. NeuroImage. 2001;14(1):21-36.
25. Szegedy C, Ioffe S, Vanhoucke V, Alemi AA, editors. Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning. National Conference on Artificial Intelligence; 2016.
26. Ellis KA, Rowe CC, Villemagne VL, Martins RN, Masters CL, Salvado O, et al. Addressing population aging and Alzheimer's disease through the Australian Imaging Biomarkers and Lifestyle study: Collaboration with the Alzheimer's Disease Neuroimaging Initiative. Alzheimer's & dementia. 2010;6(3):291-6.
27. Malone IB, Cash D, Ridgway GR, MacManus DG, Ourselin S, Fox NC, et al. MIRIAD--Public release of a multiple time point Alzheimer's MR imaging dataset. NeuroImage. 2013;70:33-6.
28. Marcus DS, Wang TH, Parker J, Csernansky JG, Morris JC, Buckner RL. Open Access Series of Imaging Studies (OASIS): cross-sectional MRI data in young, middle aged, nondemented, and demented older adults. J Cogn Neurosci. 2007;19(9):1498-507.
29. Gauthier S, Reisberg B, Zaudig M, Petersen RC, Ritchie K, Broich K, et al. Mild cognitive impairment. The lancet. 2006;367(9518):1262-70.
30. Yee E, Ma D, Popuri K, Wang L, Beg MF, The Alzheimer's Disease Neuroimaging I, et al. Construction of MRI-Based Alzheimer's Disease Score Based on Efficient 3D Convolutional Neural Network: Comprehensive Validation on 7,902 Images from a Multi-Center Dataset. J Alzheimers Dis. 2021;79(1):47-58.
31. Ansart M, Epelbaum S, Bassignana G, Bone A, Bottani S, Cattai T, et al. Predicting the progression of mild cognitive impairment using machine learning: A systematic, quantitative and critical review. Med Image Anal. 2021;67:101848.
32. Selkoe DJ. Preventing Alzheimer's disease. Science. 2012;337(6101):1488-92.
33. Frisoni GB, Fox NC, Jack CR, Jr., Scheltens P, Thompson PM. The clinical use of structural MRI in Alzheimer disease. Nat Rev Neurol. 2010;6(2):67-77.
34. Abrol A, Bhattarai M, Fedorov A, Du Y, Plis S, Calhoun V, et al. Deep residual learning for neuroimaging: An application to predict progression to Alzheimer's disease. J Neurosci Methods. 2020;339:108701.
35. Wachinger C, Salat DH, Weiner M, Reuter M, Initiative AsDN. Whole-brain analysis reveals increased neuroanatomical asymmetries in dementia for hippocampus and amygdala. Brain. 2016;139(12):3253-66.
36. Derflinger S, Sorg C, Gaser C, Myers N, Arsic M, Kurz A, et al. Grey-matter atrophy in Alzheimer's disease is asymmetric but not lateralized. Journal of Alzheimer's Disease. 2011;25(2):347-57.
37. Liu J, Wang Y, Qiao Y, editors. Sparse deep transfer learning for convolutional neural network. Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence; 2017.
38. Ke A, Ellsworth W, Banerjee O, Ng AY, Rajpurkar P. CheXtransfer: Performance and Parameter Efficiency of ImageNet Models for Chest X-Ray Interpretation. arXiv preprint arXiv:210106871. 2021.
39. Joel D, Berman Z, Tavor I, Wexler N, Gaber O, Stein Y, et al. Sex beyond the genitalia: The human brain mosaic. Proc Natl Acad Sci U S A. 2015;112(50):15468-73.
40. Eliot L, Ahmed A, Khan H, Patel J. Dump the "dimorphism": Comprehensive synthesis of human brain studies reveals few male-female differences beyond size. Neurosci Biobehav Rev. 2021;125:667-97.
41. Forest MG, Peretti ED, Bertrand J. Hypothalamic-pituitary-gonadal relationships in man from birth to puberty. Clin Endocrinol (Oxf). 1976;5(5):551-69.
42. Makris N, Swaab DF, Der Kouwe AJWV, Abbs B, Boriel D, Handa RJ, et al. Volumetric parcellation methodology of the human hypothalamus in neuroimaging: Normative data and sex differences. NeuroImage. 2013;69:1-10.
43. Raz N, Gunningdixon FM, Head D, Williamson A, Acker JD. Age and Sex Differences in the Cerebellum and the Ventral Pons: A Prospective MR Study of Healthy Adults. Am J Neuroradiol. 2001;22(6):1161-7.
44. Raz N, Dupuis JH, Briggs SD, Mcgavran C, Acker JD. Differential effects of age and sex on the cerebellar hemispheres and the vermis: a prospective MR study. Am J Neuroradiol. 1998;19(1):65-71.
45. Deng J, Dong W, Socher R, Li L-J, Li K, Fei-Fei L, editors. Imagenet: A large-scale hierarchical image database. 2009 IEEE conference on computer vision and pattern recognition; 2009: Ieee.
46. Fischl B. FreeSurfer. NeuroImage. 2012;62(2):774-81.
47. Esteban O, Markiewicz CJ, Blair RW, Moodie CA, Isik AI, Erramuzpe A, et al. fMRIPrep: a robust preprocessing pipeline for functional MRI. Nat Med. 2019;16(1):111-6.
48. Coalson TS, Van Essen DC, Glasser MF. The impact of traditional neuroimaging methods on the spatial localization of cortical areas. Proc Natl Acad Sci U S A. 2018;115(27):e6356-e65.

Figures

Figure 1

Flow diagram for the Alzheimer disease (AD) transfer learning framework and cross-validation procedure (A) Schema for the 3D Inception-ResNet-V2 model and the transfer learning framework for the Alzheimer disease classifier. (B) Schematic diagram for the leave-datasets-out 5-fold cross-validation for the sex classifier. (C) Schematic diagram for the leave-sites-out 5-fold cross-validation for the AD classifier.

Figure 2

Performance of the sex classifier (A) The receiver operating characteristic curve of the sex classifier. (B) The tensorboard monitor graph of the sex classifier in the training sample. The curve was smoothed for better visualization. (C) The tensorboard monitor graph of the sex classifier in the validation sample.

Figure 3

Performance of the Alzheimer's disease (AD) classifier Left panel shows the training and testing performance of the AD classifier on ADNI sample. Right panel shows the testing performance of the AD classifier on independent samples. (A) The receiver operating characteristic curve of the AD classifier. (B) The tensorboard monitor panel of the AD classifier in the training sample. (C) The tensorboard monitor panel of the AD classifier in the validation sample. (D) The ROC curve of AD classifier tested on the AIBL sample. (E) The ROC curve of the AD classifier tested on the MIRIAD sample. (F) The ROC curve of the AD classifier tested on the ADNI MCI sample.

Figure 4

Interpretation of the deep learning classifiers with occlusion maps Classifier performance dropped considerably when the brain areas rendered in red were masked out of the model input. (A) The occlusion maps for the sex classifier. Hypothalamus and pituitary were marked in dash line and solid line. (B) The occlusion maps for Alzheimer disease classifier.

Figure 5

Correlations between the output of the Alzheimer's disease (AD) classifier and the severity of illness The predicted scores from the AD classifier showed significant negative correlations with the mini-mental state examination (MMSE) scores of AD, normal control (NC) and mild cognitive impairment (MCI) samples. (A) Correlation between the predicted scores from AD classifier and the MMSE scores of AD samples. (B) Correlation between the predicted scores from the AD classifier and the MMSE scores of MCI samples. (C) Correlation between the predicted scores from AD classifier and the MMSE scores of NC samples. (D) Correlation between the predicted scores from AD classifier and the MMSE scores of AD, NC, and MCI samples.

Supplementary Files

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

- LuSupplementaryMaterials.docx