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

Bin Lu[1,2], Hui-Xian Li[1,2], Zhi-Kai Chang[1,2], Le Li[3], Ning-Xuan Chen[1,2], Zhi-Chen Zhu[1,2], Hui-Xia Zhou[1,2], Xue-Ying Li[1,4,5], Yu-Wei Wang[1,2], Shi-Xian Cui[1,4,5], Zhao-Yu Deng[1,2], Zhen Fan[6], Hong Yang[7], Xiao Chen[1,2], Paul M. Thompson[8], Francisco Xavier Castellanos[9,10], Chao-Gan Yan[1,2,11,12*], for the Alzheimer's Disease Neuroimaging Initiative[**]

[1]CAS Key Laboratory of Behavioral Science, Institute of Psychology, Beijing, China; [2]Department of Psychology, University of Chinese Academy of Sciences, Beijing, China; [3]Center for Cognitive Science of Language, Beijing Language and Culture University, Beijing, China; [4]Sino-Danish College, University of Chinese Academy of Science, Beijing, China; [5]Sino-Danish Center for Education and Research, Beijing, China; [6]Department of Neurosurgery, Huashan Hospital, Fudan University, Shanghai Neurosurgical Clinical Center, Shanghai, China; [7]Department of Radiology, The First Affiliated Hospital, College of Medicine, Zhejiang University, Hangzhou, Zhejiang, China; [8]Imaging Genetics Center, Mark & Mary Stevens Institute for Neuroimaging & Informatics, Keck School of Medicine, University of Southern California, Los Angeles, CA, USA; [9]Department of Child and Adolescent Psychiatry, NYU Grossman School of Medicine, New York, NY, USA; [10]Nathan Kline Institute for Psychiatric Research, Orangeburg, NY, USA; [11]International Big-Data Research Center for Depression (IBRCD), Institute of Psychology, Chinese Academy of Sciences, Beijing, China; [12]Magnetic Resonance Imaging Research Center, Institute of Psychology, Chinese Academy of Sciences, Beijing, China. *e-mail: ycg.yan@gmail.com. **Data used in 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 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

Abstract

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. Here we aimed to build a practical brain imaging-based AD diagnostic classifier using deep learning/transfer learning on dataset of unprecedented size and diversity. We pooled MRI data from more than 217 sites/scanners to constitute the largest brain MRI sample to date (85,721 scans from 50,876 participants). Next, we applied a state-of-the-art deep convolutional neural network, Inception-ResNet-V2, to build 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. 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) dataset and 94.2%/87.9% accuracy for direct tests on two unseen independent datasets (AIBL/OASIS). 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. In sum, the proposed AD classifier could offer a medical-grade biomarker that could be integrated into AD diagnostic practice. Our trained model, code and preprocessed data are freely available to the research community.

Keywords

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

1. 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, biomarkers such as amyloid measures from cerebrospinal fluid (CSF) and amyloid-sensitive positron emission tomography (PET) have been integrated to improve the specificity in diagnosing AD[1,2]. However, the invasive nature and the low sensitivity of these biomarkers hampers their application in routine clinical settings. For example, the IWG-1 criteria only achieved 68% sensitivity and the NIA-AA criteria achieved only 65.6% sensitivity[3,4]. A novel non-invasive biomarker with both high sensitivity and specificity is needed for diagnosing AD. Structural MRI is a promising candidate considering its non-invasive nature and wider availability than PET. In addition, there are well-developed MRI data preprocessing pipelines that make it feasible to integrate MRI biomarkers 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[5], aftershock pattern prediction[6] and automatic speech recognition[7]. 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[8] and skin cancer[9], and for breast cancer screening[10]. Given the success of recent CNN algorithms in pattern recognition, similar MRI-based diagnostic classifiers are likely to be highly valuable, if they can be integrated into routine neurological practice.

Even so, prior attempts at MRI-based AD diagnosis have yet to reach 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 and magnetic field strength, head coil hardware, the pulse sequence, applied gradient fields, reconstruction methods, scanning parameters, voxel size, field of view, etc. Participants also differ in sex, age, race and education, and 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. For instance, Qiu and colleagues built a deep-learning classifier for AD with an average accuracy of 82.2% using brain imaging data from four datasets[11]. However, when tested on the FHS (Framingham Heart Study) dataset, the accuracy and specificity of the AD classifier dropped to 76.6% and 71.2%, with a relatively high sensitivity (90.1%). On the contrary, when tested on AIBL dataset (from the Australian Imaging, Biomarker and Lifestyle Study of Ageing), the same classifier achieved relatively high accuracy and specificity (87.0% and 92.4%), but the sensitivity was poor (59.4%). The variable accuracy and inconsistent tradeoff between sensitivity and specificity in data from different medical centers hampers these proposed methods from being deployed across multiple clinical institutions. To alleviate the unsatisfactory generalization performance, Bashyam et al. used a more heterogeneous sample to build a brain age prediction model that would be more generalizable to data from unseen sites/scanners[12]. However, when transfer learning to AD, they only used random cross-validation on the ADNI dataset with an accuracy of 86%, and did not implement independent dataset validation. Reviews of brain imaging-based AD classifiers suggest that most machine learning methods have been trained on scans from only a few hundreds of participants, which makes them unable to achieve stable performance when validated independently[13].

Therefore, one bottleneck in developing a practical brain imaging-based classifier is the variety and comprehensiveness of training datasets. As most publicly available AD datasets only contain several hundred patients and corresponding healthy controls, directly training models on these datasets may result in overfitting with poor generalization to unseen test data[13]. In the current paper, we propose a transfer learning framework to solve this problem, by training a model on a certain characteristic for which there are abundant samples available, and fine-tuning it to another characteristic in smaller samples[14], following successful examples in diagnosing retinal disease[8] and skin cancer[9]. In the brain imaging field, the scientific community has shared hundreds of thousands of brain images from hundreds of sites/scanners all over the world. Nonetheless, no studies have fully implemented this abundant resource 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 Supplementary Table 1) to pre-train a brain imaging based classifier, in order to ensure high generalization capability. After that, the pre-trained model was fine-tuned for AD classification and was validated through leave-sites-out cross-validation and independent validation. Mild cognitive impairment (MCI) is a syndrome defined as relative cognitive decline without symptoms interfering with daily life, but more than half of MCI patients progress to dementia within 5 years[15]. Discrimination between MCI patients who will progress to AD (pMCI) and MCI patients who will not progress to AD over a given time interval (stable MCI, or sMCI) would facilitate the early treatment of pMCI. Therefore, we used the present AD classifier to predict progression in MCI patients to further evaluate its generalization capability.

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 datasets 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 validation was 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 framework, and have built an online predicting website (http://brainimagenet.org) for anyone interested in testing our classifier with brain imaging data from any research participants and any scanner.

2. Materials and 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 S1**. 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.

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)[16], which is based on SPM[17], 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[18]. Two voxel-based structural metrics, GMD and GMV were fed into the deep learning classifier as two features for each participant. GMV was based on modulated GMD images by using the Jacobian determinants derived from the spatial normalization in the VBM analysis[19].

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 (See **Fig. S6** for the distribution of correlations for each dataset). 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.

Deep learning: classifier training and testing for sex. We trained a 3-dimensional Inception-ResNet-v2[20] model adopted from its 2-dimensional version in the Keras built-in application (see **Fig. 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 **Fig. 1B** 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.

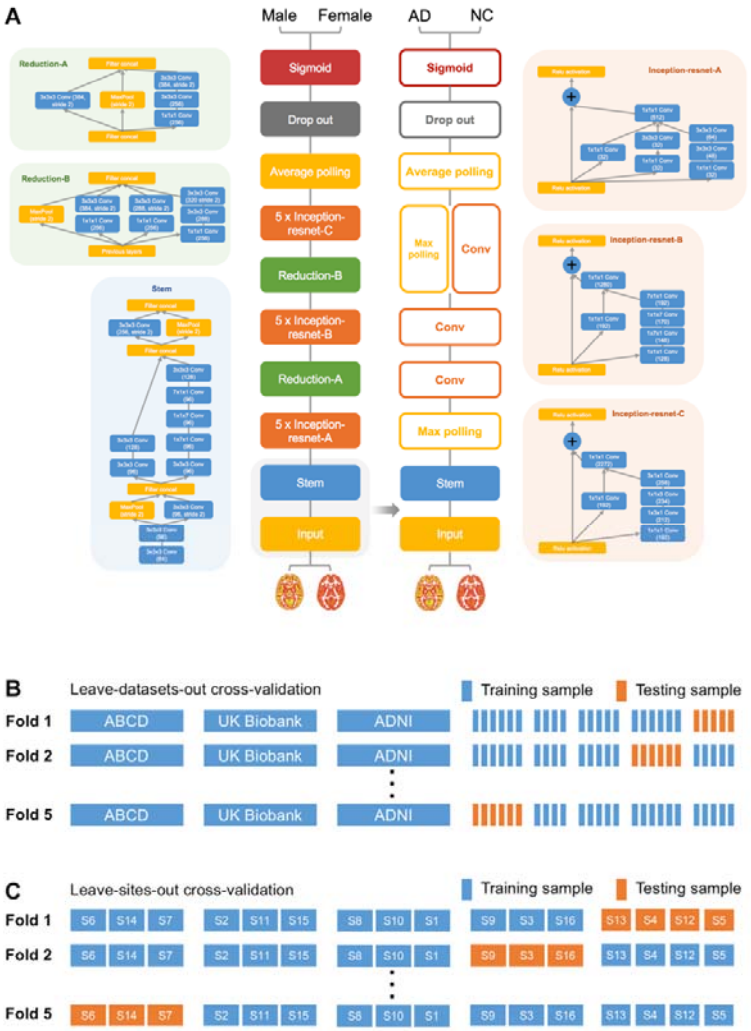

Fig. 1 | Flow diagram for the Alzheimer disease (AD) transfer learning framework and cross-validation procedure. (**A**) Schema for the 3D Inception-ResNet-V2 model and the transfer learning framework for the Alzheimer disease classifier. (**B**) Schematic diagram for the leave-datasets-out 5-fold cross-validation for the sex classifier. (**C**) Schematic diagram for the leave-sites-out 5-fold cross-validation for the AD classifier.

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)). 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 **Fig. 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.

Furthermore, to test the generalizability of the AD classifier, we directly tested the classifier on another unseen independent AD sample, i.e., Australian Imaging, Biomarker and Lifestyle Flagship Study of Ageing (AIBL)[21] and Open Access Series of Imaging Studies (OASIS)[22,23]. We used the averaged output of 5 AD classifiers in the five-fold cross-validation trained on ADNI as the final output for a participant. We used diagnoses provided by the AIBL dataset as the labels of samples (101 samples from 82 AD patients and 523 samples from 324 NCs). As OASIS did not specify the criteria for an AD diagnosis, we adopted 2 criteria from ADNI-1 to define AD patients, i.e., 1) mini-mental state examination score between 20 and 26 (inclusive) and 2) clinical dementia rating score = 0.5 or 1.0. Thus, we tested on 277 AD samples and 995 normal control samples who met the ADNI-1 criteria for AD and NC in the OASIS dataset. Of note, AIBL and OASIS scanning conditions and recruitment criteria

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 MCI. MCI is a diagnosis defined as cognitive decline without impairment in everyday activities[15]. People with the amnestic subtype of MCI have a high risk of converting to AD. 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' (i.e., their early scans before entering the AD phase; the images labeled 'Conversion: MCI to AD' and images labeled as 'AD' after conversion were not included). 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 mini-mental state examination (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.

3. 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 S1**; there were no application requirements for some datasets). Data for each participant contained at least one session with a T1-weighted brain structural image and information on participant sex. After quality control, all these samples were used to pre-train the stem module to achieve better generalization for further AD classifier training. For the further fine-tuning of the AD classifier, ADNI (16,596 samples from 2,212 participants), AIBL (624 samples from 406 participants) and OASIS (3,150 samples from 1,664 participants) 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 Fig. 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. The brain structural data were segmented and normalized to yield grey matter density (GMD) and grey matter volume (GMV) maps (i.e., GMD and GMV maps were treated as different input channels). 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 (see **Fig. 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).

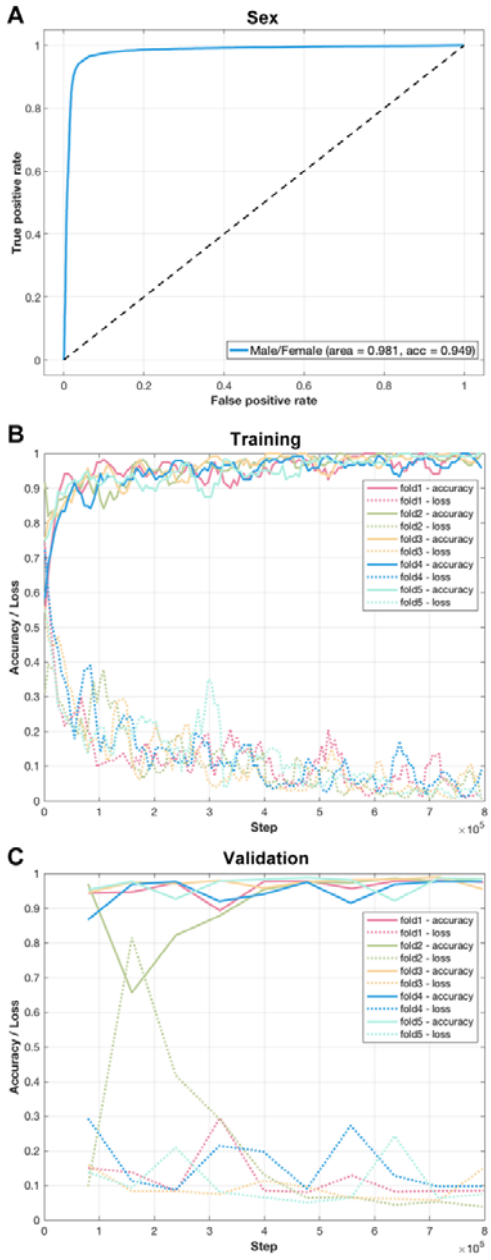

Fig. 2 | Performance of the sex classifier. (**A**) The receiver operating characteristic curve of the sex classifier. (**B**) The tensorboard monitor graph of the sex classifier in the training sample. The curve was smoothed for better visualization. (**C**) The tensorboard monitor graph of the sex classifier in the validation sample.

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 **Fig. 3** and **Table 1**). 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, see **Fig. S7**).

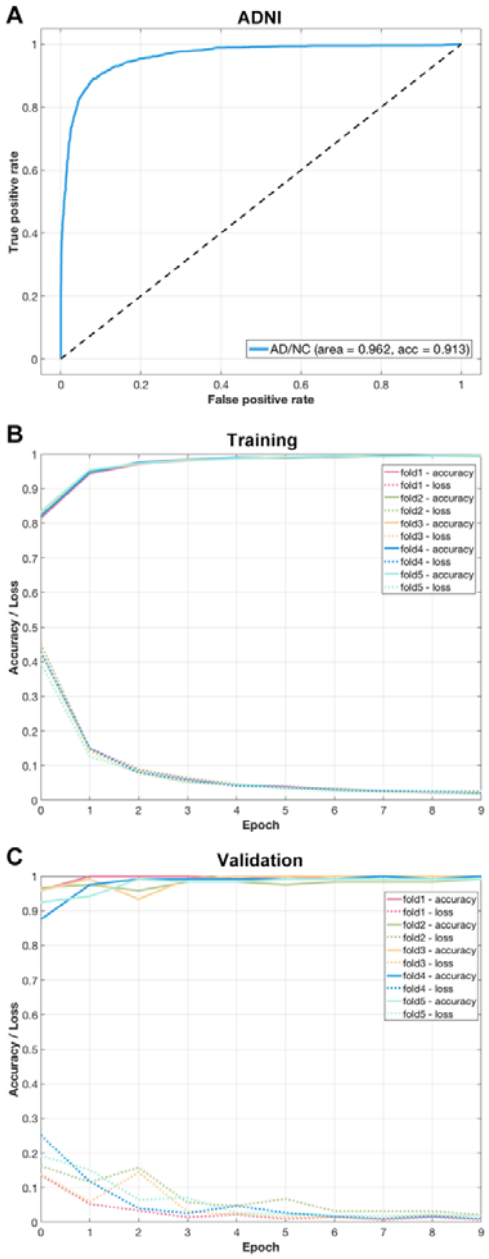

Fig. 3 | Performance of the Alzheimer's disease (AD) classifier. (A) The receiver operating characteristic curve of the AD classifier. **(B)** The tensorboard monitor panel of the AD classifier in the training sample. **(C)** The tensorboard monitor panel of the AD classifier in the validation sample.

To test the generalizability of the AD classifier, we applied it to unseen independent AD datasets, i.e., AIBL and OASIS 1 and 2. The AD classifier achieved 94.2% accuracy in AIBL

with 0.97 AUC (see **Fig. 4A**). Sensitivity and specificity were 0.881 and 0.954, respectively. The AD classifier achieved 87.9% accuracy in OASIS with 0.936 AUC (see **Fig. 4B**). Sensitivity and specificity were 0.796 and 0.902, respectively.

Table 1 | Performance of the Alzheimer's disease classifier

Dataset	n (AD)	n (NC)	Accuracy	AUC	Sensitivity	Specificity
ADNI	2,186	4,671	0.913	0.962	0.848	0.943
AIBL	101	523	0.942	0.97	0.881	0.954
OASIS	277	995	0.879	0.936	0.796	0.902

AD = Alzheimer's disease; NC = normal control. The sample sizes shown here are the numbers of T1-weighted brain MRI scans.

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. 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.

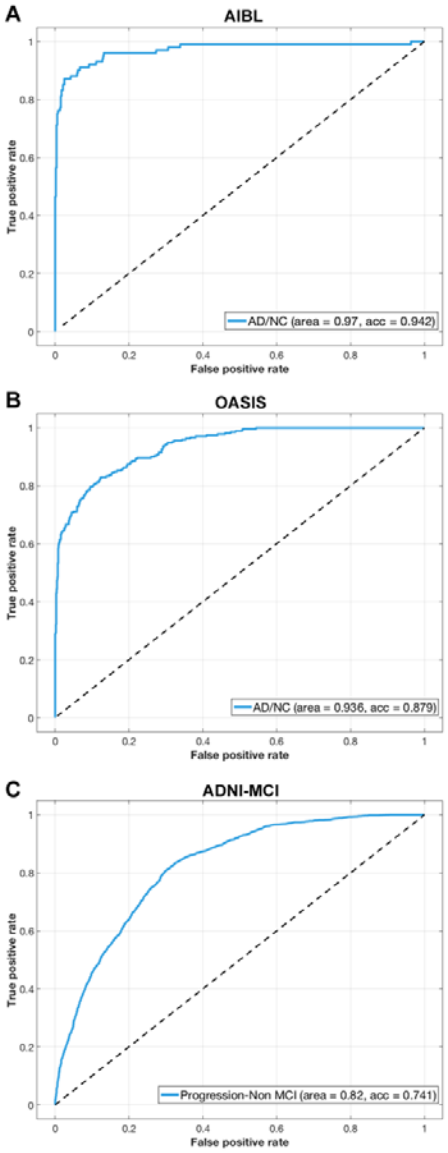

Fig. 4 | Receiver operating characteristic (ROC) curves for the Alzheimer disease (AD) classifier when tested on independent AD samples and a mild cognitive impairment sample. (**A**) The ROC curve of AD classifier tested on the AIBL sample. (**B**) The ROC curve of AD classifier tested on the OASIS sample. (**C**) The ROC curve of AD classifier tested on MCI sample in ADNI. The images of MCI subjects with future conversion to AD were labeled as "AD", and the images of MCI subjects who had not shown conversion to AD were labeled as "NC".

As a supplementary analysis, we also trained the AD classifier that kept the intact structure of the base model in transfer learning. 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 (see **Fig. S2A**). 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 (see **Fig. S3A**). Sensitivity and specificity were 0.851 and 0.924, respectively. The AD classifier achieved 86.1% accuracy in OASIS with 0.921 AUC (see Fig. S3B). 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 (see **Fig. S3C**).

3.5. Interpretation of the deep learning classifiers

To better understand the brain imaging-based deep learning classifier, we calculated occlusion maps for the classifiers. In brief, we continuously set a cubic brain area of every input image to zeros, and made the classifier attempt classification based on the defective samples. The occlusion map showed that hypothalamus, superior vermis, thalamus, amygdala, putamen, accumbens, hippocampus and parahippocampal gyrus played critical roles in predicting sex (see **Fig. 5A**). 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 (see **Fig. 5B**).

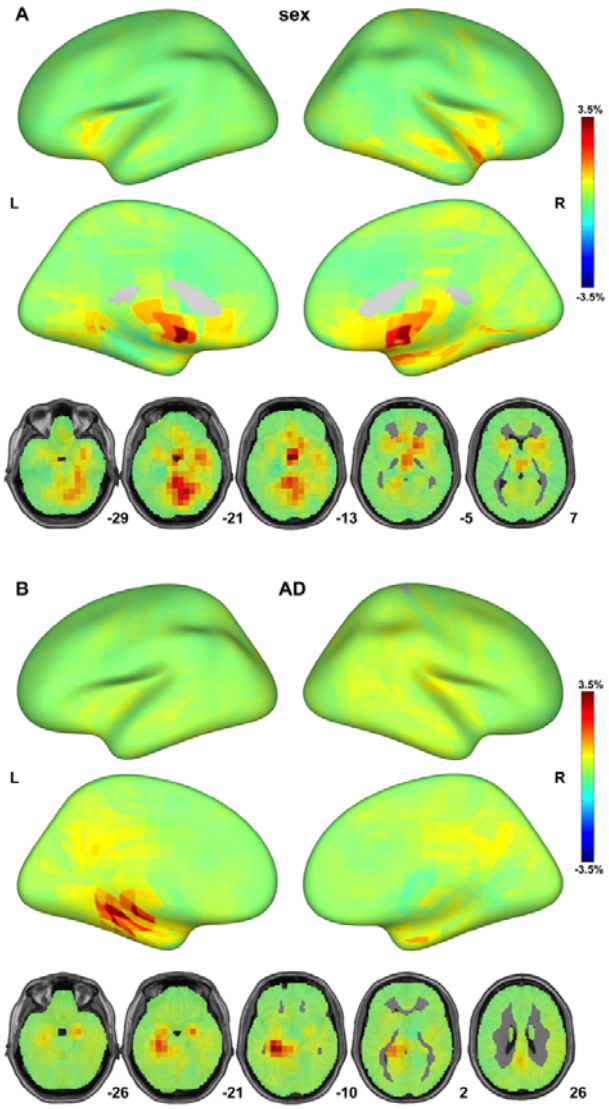

Fig. 5 | Interpretation of the deep learning classifiers with occlusion maps. Classifier performance dropped considerably when the brain areas rendered in red were masked out of the model input. (**A**) The occlusion maps for the sex classifier. (**B**) The occlusion maps for Alzheimer disease classifier.

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 mini-mental state examination (MMSE, provided by ADNI datasets) 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, $p < 1 \times 10^{-307}$) (See Fig. 6). 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, See **Fig. S5**). Importantly, the effect size of the predicted scores by the classifier is much larger than the behavioral measure (MMSE scores).

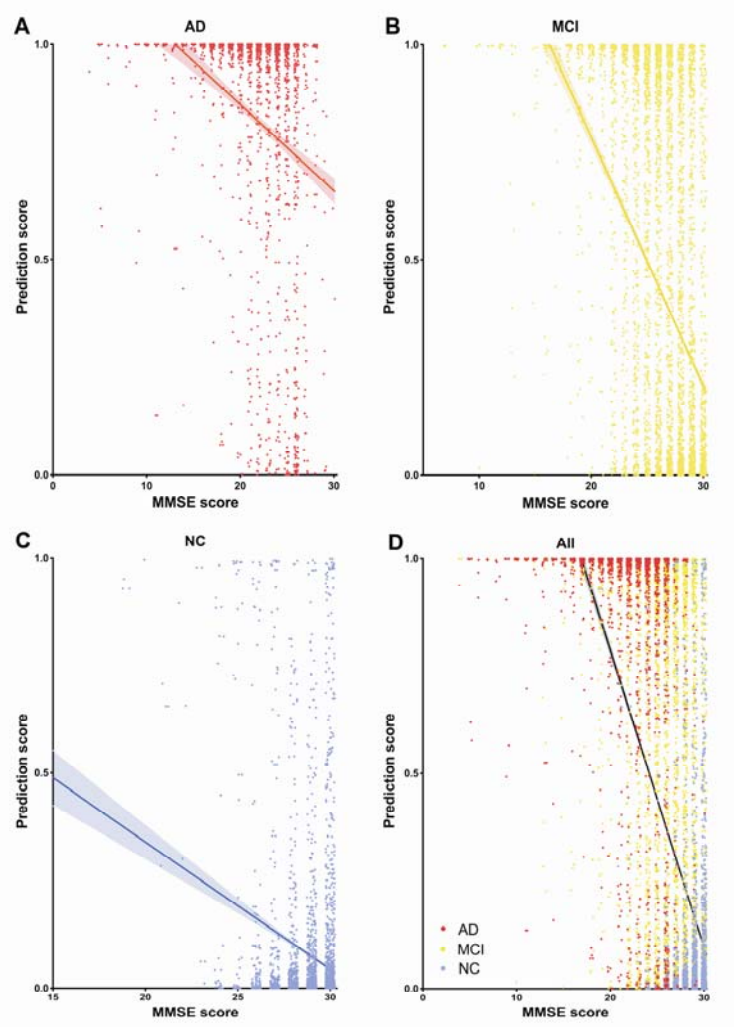

Fig. 6 | Correlations between the output of the Alzheimer's disease (AD) classifier and the severity of illness. The predicted scores from the AD classifier showed significant negative correlations with the mini-mental state examination (MMSE) scores of AD, normal

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.

4. Discussion

Using an unprecedentedly diverse brain imaging sample, we pre-trained an industrial-grade 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%/87.9% 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 industrial-grade high accuracy and generalization capability of our deep neural network classifiers demonstrate that brain imaging did have practical utility for auxiliary diagnosis. The current prototype may facilitate future research to apply brain imaging in many practical application fields. Of note, the output of the deep neural network model is a continuous variable, so the threshold can be adjusted to balance sensitivity and specificity. For example, when tested on the independent sample (OASIS), sensitivity and specificity results were 0.796 and 0.902, 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 0.881 at a cost of decreasing specificity to 0.796. Thus, in our freely available AD prediction website, users can obtain continuous outputs and adjust the threshold by themselves. This adjustable characteristic of the present model makes itself more suitable for integration into the current

diagnostic criteria as a diagnostic biomarker. The proposed MRI-based biomarker has a high sensitivity, which may address the lack of sensitivity of other biomarkers.

Except for the feasibility of being integrated into diagnostic criteria, the present AD model also showed outstanding characteristics as a progression biomarker. 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[24]. 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[13,24]. 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[25]. Clinicians can use the present AD classifier as auxiliary decision support system. Second, the output of the deep neural network can indicate the clinical severity for AD patients or people with MCI, as the predicted scores showed significant negative correlations with MMSE scores. Considering the "greedy" characteristic of CNN for reducing training loss, the prediction scores for AD and NC were overstated, so the magnitude of negative correlations might be even underestimated. Third, when directly comparing the predicted scores (or MMSE scores) between sMCI and pMCI groups, the predicted scores showed much higher effect size than MMSE scores (Cohen's $d_{prediction}$ = 1.08 vs Cohen's d_{MMSE} = -0.73), indicating that predicted scores may offer better prompting or 'warning' effects for the physician to differentiate MCI patients.

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 biomarker was in line with former pathological findings and clinical practices. For example, AD induced

457 brain structural changes have been frequently reported by MRI studies. Among all the
458 structural findings, hippocampal atrophy is the most prominent change and is used in imaging
459 assisted diagnosis[26]. Neurobiological changes in the hippocampus typically precede
460 progressive neocortical damage and AD symptoms. The convergence of our deep learning
461 system and human physicians on alterations in hippocampal structure for classifying AD
462 patients is in line with the crucial role of the hippocampus in AD. Furthermore, brain atrophy
463 in AD has been frequently reported as left lateralized[27,28]. Compared to the un-optimized AD
464 classifier, a slight left hemisphere preference for input features may help explain the
465 improved performance of the optimized AD classifier (see **Fig. 5** and **Fig. S4** to compare the
466 occlusion maps).

467

468 Rather than indiscriminately imitate the structure of the base model in transfer learning, the
469 present AD classifier significantly simplified the model before the fine-tuning procedure. In
470 fact, the performance of the unoptimized AD classifier was far poorer than that of the
471 optimized AD classifier in accuracy, sensitivity, specificity, and in independent validation
472 performance (see **Fig. S2-3**). There is some reported evidence that truncating or pruning
473 models before transfer learning may facilitate the performance of the transferred models[29,30].
474 As the sample for training the AD classifier is considerably smaller than that used to train the
475 sex classifier, the simplified model structure may help to avoid overfitting and improve
476 generalizability.

477

478 By precisely predicting the sex of people, the present study also advances our understanding
479 of sex differences in human brain. Daphna and colleagues extracted hundreds of VBM
480 features from structural MRI and concluded that “the so-called male/female brain” does not
481 exist as no individual structural feature supports a sexually dimorphic view of human brains[31].
482 However, human brains may embody sexually dimorphic features in a multivariate manner.
483 The high accuracy and generalizability of the present sex classifier demonstrate that sex is
484 separable in a 1,981,440-dimension (96*120*86*2) feature space. Among those 1,981,440
485 features, hypothalamus 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[32]. Men have significantly larger hypothalamus than women relative to cerebrum size[33]. 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[34]. 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[35]. 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 (https://github.com/Chaogan-Yan/BrainImageNet), 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[36], fmriprep[37] and DPABISurf, surface-based algorithms have shown their superiority when compared with traditional volume-based algorithms[38]. 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 biomarkers.

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 an industrial-grade 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.

Data and code availability

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.

Acknowledgement

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.

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).

Competing interests

The authors declare no competing interests.

Supplementary material

Supplementary material is available online.

References

1 Dubois, B. *et al.* Advancing research diagnostic criteria for Alzheimer's disease: the IWG-2 criteria. *The Lancet Neurology* **13**, 614-629, (2014).

2 Jack Jr, C. R. *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* **7**, 257-262, (2011).

3 de Jager, C. A., Honey, T. E., Birks, J. & Wilcock, G. K. Retrospective evaluation of revised criteria for the diagnosis of Alzheimer's disease using a cohort with post-mortem diagnosis. *Int. J. Geriatr. Psychiatry* **25**, 988-997, (2010).

4 Harris, J. M. *et al.* Do NIA-AA criteria distinguish Alzheimer's disease from frontotemporal dementia? *Alzheimer's & Dementia* **11**, 207-215, (2015).

5 Ham, Y.-G., Kim, J.-H. & Luo, J.-J. Deep learning for multi-year ENSO forecasts. *Nature* **573**, 568-572, (2019).

6 DeVries, P. M. R., Viegas, F., Wattenberg, M. & Meade, B. J. Deep learning of aftershock patterns following large earthquakes. *Nature* **560**, 632-634, (2018).

7 Liu, W. *et al.* A survey of deep neural network architectures and their applications. *Neurocomputing* **234**, 11-26, (2017).

8 Kermany, D. S. *et al.* Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell* **172**, 1122-1131, (2018).

9 Esteva, A. *et al.* Dermatologist-level classification of skin cancer with deep neural networks. *Nature* **542**, 115-118, (2017).

10 McKinney, S. M. *et al.* International evaluation of an AI system for breast cancer screening. *Nature* **577**, 89-94, (2020).

11 Qiu, S. *et al.* Development and validation of an interpretable deep learning framework for Alzheimer's disease classification. *Brain* **143**, 1920-1933, (2020).

12 Bashyam, V. 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* **143**, 2312-2324, (2020).

13 Rathore, S., Habes, M., Iftikhar, M. A., Shacklett, A. & Davatzikos, C. A review on neuroimaging-based classification studies and associated feature extraction methods for Alzheimer's disease and its prodromal stages. *NeuroImage* **155**, 530-548, (2017).

14 Yosinski, J., Clune, J., Bengio, Y. & Lipson, H. in *Adv. Neural Inf. Process. Syst.* 3320-3328.

15 Gauthier, S. *et al.* Mild cognitive impairment. *The lancet* **367**, 1262-1270, (2006).

16 Yan, C. G. & Zang, Y. F. DPARSF: A MATLAB Toolbox for "Pipeline" Data Analysis of Resting-State fMRI. *Front. Syst. Neurosci.* **4**, 13, (2010).

17 Friston, K. J. *et al.* Statistical parametric maps in functional imaging: a general linear approach. *Hum. Brain Mapp.* **2**, 189-210, (1994).

18 Goto, M. *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* **55**, 869-875, (2013).

19 Good, C. D. *et al.* A voxel-based morphometric study of ageing in 465 normal adult human brains. *NeuroImage* **14**, 21-36, (2001).

20 Szegedy, C., Ioffe, S., Vanhoucke, V. & Alemi, A. A. in *National Conference on Artificial Intelligence.* 4278-4284.

21 Ellis, K. A. *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* **6**, 291-296, (2010).

22 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).

23 Marcus, D. S., Fotenos, A. F., Csernansky, J. G., Morris, J. C. & Buckner, R. L. Open access series of

imaging studies: longitudinal MRI data in nondemented and demented older adults. *J. Cogn. Neurosci.* **22**, 2677-2684, (2010).

24 Ansart, M. *et al.* Predicting the progression of mild cognitive impairment using machine learning: A systematic, quantitative and critical review. *Med. Image Anal.* **67**, 101848, (2021).

25 Selkoe, D. J. Preventing Alzheimer's disease. *Science* **337**, 1488-1492, (2012).

26 Frisoni, G. B., Fox, N. C., Jack, C. R., Jr., Scheltens, P. & Thompson, P. M. The clinical use of structural MRI in Alzheimer disease. *Nat. Rev. Neurol.* **6**, 67-77, (2010).

27 Wachinger, C., Salat, D. H., Weiner, M., Reuter, M. & Initiative, A. s. D. N. Whole-brain analysis reveals increased neuroanatomical asymmetries in dementia for hippocampus and amygdala. *Brain* **139**, 3253-3266, (2016).

28 Derflinger, S. *et al.* Grey-matter atrophy in Alzheimer's disease is asymmetric but not lateralized. *Journal of Alzheimer's Disease* **25**, 347-357, (2011).

29 Liu, J., Wang, Y. & Qiao, Y. in *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*. 2245-2251.

30 Ke, A., Ellsworth, W., Banerjee, O., Ng, A. Y. & Rajpurkar, P. CheXtransfer: Performance and Parameter Efficiency of ImageNet Models for Chest X-Ray Interpretation. *arXiv preprint arXiv:2101.06871*, (2021).

31 Joel, D. *et al.* Sex beyond the genitalia: The human brain mosaic. *Proc. Natl. Acad. Sci. U. S. A.* **112**, 15468-15473, (2015).

32 Forest, M. G., Peretti, E. D. & Bertrand, J. Hypothalamic-pituitary-gonadal relationships in man from birth to puberty. *Clin. Endocrinol. (Oxf.)* **5**, 551-569, (1976).

33 Makris, N. *et al.* Volumetric parcellation methodology of the human hypothalamus in neuroimaging: Normative data and sex differences. *NeuroImage* **69**, 1-10, (2013).

34 Deng, J. *et al.* in *2009 IEEE conference on computer vision and pattern recognition.* 248-255 (Ieee).

35 Hendrycks, D., Lee, K. & Mazeika, M. Using pre-training can improve model robustness and uncertainty. *arXiv preprint arXiv:1901.09960*, (2019).

36 Fischl, B. FreeSurfer. *NeuroImage* **62**, 774-781, (2012).

37 Esteban, O. *et al.* fMRIPrep: a robust preprocessing pipeline for functional MRI. *Nat. Med.* **16**, 111-116, (2019).

38 Coalson, T. S., Van Essen, D. C. & Glasser, M. F. The impact of traditional neuroimaging methods on the spatial localization of cortical areas. *Proc. Natl. Acad. Sci. U. S. A.* **115**, e6356-e6365, (2018).