Brain Disorders 1 (2021) 100005

Contents lists available at ScienceDirect

Brain Disorders

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

AI4AD: Artificial intelligence analysis for Alzheimer's disease classification based on a multisite DTI database

Yida Qu[a,b], Pan Wang[c], Bing Liu[a,b,p], Chengyuan Song[d], Dawei Wang[e], Hongwei Yang[f], Zengqiang Zhang[g], Pindong Chen[a,b], Xiaopeng Kang[a,b], Kai Du[a,b], Hongxiang Yao[h], Bo Zhou[i], Tong Han[j], Nianming Zuo[a,b], Ying Han[k,m,n,o], Jie Lu[f], Chunshui Yu[l], Xi Zhang[i], Tianzi Jiang[a,b,p], Yuying Zhou[c], Yong Liu[a,b,p,q,*], Multi-Center Alzheimer's Disease Imaging (MCADI) Consortium

[a] *Brainnetome Center & National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China*
[b] *University of Chinese Academy of Sciences, Beijing, China*
[c] *Department of Neurology, Tianjin Huanhu Hospital, Tianjin University, Tianjin, China*
[d] *Department of Neurology, Qilu Hospital of Shandong University, Ji'nan, China*
[e] *Department of Radiology, Qilu Hospital of Shandong University, Ji'nan, China*
[f] *Department of Radiology, Xuanwu Hospital of Capital Medical University, Beijing, China*
[g] *Branch of Chinese PLA General Hospital, Sanya, China*
[h] *Department of Radiology, the Second Medical Centre, National Clinical Research Centre for Geriatric Diseases, Chinese PLA General Hospital, Beijing, China*
[i] *Department of Neurology, the Second Medical Centre, National Clinical Research Centre for Geriatric Diseases, Chinese PLA General Hospital, Beijing, China*
[j] *Department of Radiology, Tianjin Huanhu Hospital, Tianjin, China*
[k] *Department of Neurology, Xuanwu Hospital of Capital Medical University, Beijing, China*
[l] *Department of Radiology, Tianjin Medical University General Hospital, Tianjin, China*
[m] *Beijing Institute of Geriatrics, Beijing, China*
[n] *National Clinical Research Center for Geriatric Disorders, Beijing, China*
[o] *Center of Alzheimer's Disease, Beijing Institute for Brain Disorders, Beijing, China*
[p] *Center for Excellence in Brain Science and Intelligence Technology, Institute of Automation, Chinese Academy of Sciences, Beijing, China*
[q] *School of Artificial Intelligence, Beijing University of Posts and Telecommunications, Beijing, China*

ARTICLE INFO

Keywords:
Alzheimer's disease (AD)
Diffusion tensor imaging (DTI)
Multisite
Automated fiber quantification (AFQ)
Classification

ABSTRACT

Background: Diffusion tensor imaging (DTI) has been widely used to identify structural integrity and to delineate white matter (WM) degeneration in Alzheimer's disease (AD). However, the validity and replicability of the ability to discriminate AD patients and normal controls (NCs) of WM measures are limited due to the use of small cohorts and diverse image processing methods. As yet, we still do not have a clear idea of whether WM characteristics are biomarkers for AD.

Methods: We conducted a competition with diffusion measurements along 18 fiber tracts as features extracted via the automated fiber quantification (AFQ) method based on one of the largest worldwide DTI multisite biobanks (862 individuals, consisting of 279 NCs, 318 ADs, and 265 MCIs). After quality control, 825 subjects (276 NCs, 294 ADs, and 255 MCIs) were divided into a public training set (N=700) and a private testing set (N=125). Forty-eight teams submitted 130 solutions that were estimated on the private testing samples. We reported the final results of the top ten models.

Results: The performance of white matter features in AD classification was stable and generalizable, which indicated the potential of WM to be a biomarker for AD. The best model achieved a prediction accuracy of 82.35% (with a sensitivity of 86.36% and a specificity of 78.05%) on the private testing set. The average accuracy of the top ten solutions was over 80%.

* Corresponding authors at: School of Artificial Intelligence, Beijing University of Posts and Telecommunications, Beijing, 100876 China.
E-mail addresses: yongliu@bupt.edu.cn, yliu@nlpr.ia.ac.cn (Y. Liu).

https://doi.org/10.1016/j.dscb.2021.100005
Received 24 December 2020; Received in revised form 26 January 2021; Accepted 5 February 2021
Available online 10 February 2021

Conclusions: The results of this competition demonstrated that DTI is a powerful tool to identify AD. A larger dataset and additional independent cohort cross-validation may improve the discriminant performance and generalization power of the classification models, thus revealing more precise disease severity factors associated with AD. For this purpose, we have released this database (https://github.com/YongLiuLab/AI4AD_AFQ) to the community, with the expectation of new solutions for the accurate diagnosis of AD.

1. Introduction

Alzheimer's disease (AD) is a chronic, progressive neurodegenerative disease that is associated with cognitive dysfunction, psychiatric symptoms, and daily life disorders that seriously threaten the normal life of patients and place a significant burden on both society and the family [1,2]. The prevalence of AD could increase in parallel with the progressive aging of the population. Researchers have conducted numerous studies on the identification of the pathogenic mechanism of this disorder, and they have also been searching for methods for the diagnosis of AD, thus concluding that an early diagnosis and investigation would have a significant positive impact on disease control. Specifically, early diagnosis enables timely disease interventions (including pharmacological symptomatic treatments and psychosocial support etc.) that might slow down the progression of the disease, which makes the early diagnosis of AD a research priority [3–5].

Considerable evidence has suggested that gray matter atrophy (as revealed via the use of structural MRI) is one of the core biomarkers of AD pathology [see reviews by 6,7]. Furthermore, there is growing interest in the application of artificial intelligence methods to functional alterations (as measured by the use of fMRI) [8,9] and the clinical application of the cerebral amyloid burden through the use of PET [10,11]. Some competitions have demonstrated the effectiveness of structural and functional changes in AD classification [12–15]. For example, the best performance has been shown to have an approximate accuracy of 85% with features based on multimodality images for the Alzheimer's Disease Prediction Of Longitudinal Evolution (TADPOLE) Challenge (https://tadpole.grand-challenge.org/) [13]. However, white matter (WM) impairment has been neglected as being an important part of the pathological cascade in AD [6,16–18]. In terms of WM, the most powerful tool is diffusion tensor imaging (DTI), which is a noninvasive in vivo imaging technique that can demonstrate structural integrity and delineate white matter degeneration in AD through diffusion properties that can identify WM fiber trends and the degree of damage [18–21]. Furthermore, several previous studies have indicated that white matter alterations may be useful for the early diagnosis of AD [22–27]. For example, in the AD versus NC binary classification task, Dyrba and colleagues (2013) used whole brain white matter measures as characteristics and SVM and naive Bayes as classifiers, which achieved an accuracy of approximately 80% through the use of SVM with pooled cross-validation [23]. Ebadi and colleagues (2017) used brain structural network graph metrics as characteristics and SVM as a classifier, with a mean accuracy of ~80% [28]. Dou and colleagues (2020) explored the classification usability of fiber bundle features that were extracted via automated fiber quantification (AFQ) with a mean accuracy of ~78% [17], and the present competition is the continuation of this study. The prediction of AD via brain diffusion imaging will provide additional biomarkers and will reveal the mechanisms of the pathology of AD.

It is very important to use large datasets, unified image processing pipelines, and independent data sets for cross-validation in the search for biomarkers for AD. Specifically, studies based on small cohorts with a single site are often unconvincing because of the insufficiency of representativeness for the entire population and incredible accuracy that is caused by possible overfitting (due to the absence of independent external validation). Furthermore, it is difficult to compare the various studies, due to site differences that are caused by machine manufacturers, acquisition parameters, and diverse image processing pipelines. [29–31]. Consequently, the use of multisite datasets with the same image processing pipeline can not only reduce confusion variance and improve the clinical representation, but can also enable independent site cross-validation, which is beneficial for reducing overfitting.

Based on the previously mentioned considerations, we conducted a competition with WM diffusion measurements to search for the best models for the potential early diagnosis of AD. To reduce the variance of preprocessing steps, as well as to meet the rapid and convenient requirement for large data sets, we selected the AFQ method to extract white matter attributes because of its advantages in overcoming the inconvenience of the hand-drawn region of interest and in automatically and efficiently detecting detailed information along 18 main fiber bundles [32,33]. This method has been widely used in brain development and aging [34], as well as for several diseases, including AD [17,27,35–37]. To test the generalization abilities of the models, we established one of the largest multisite DTI biobanks (containing 862 individuals, including 279 NCs, 318 ADs, and 265 MCIs). After quality control, 825 subjects (276 NCs, 294 ADs, and 255 MCIs) were divided into the public training set (N=700) and the private testing set (N=125) for model evaluation. All of the features and the codes of the top ten solutions are available at GitHub (https://github.com/YongLiuLab/AI4AD_AFQ).

2. Methods

2.1. Dataset and extraction of white matter features

The present dataset combined data from 7 MRI scanners in 4 hospitals in China, which contained a total of 862 individuals (279 NCs, 318 ADs, and 265 MCIs) with DTI images, T1 images, and demographic and psychological information. Detailed information can be found in our previous study [17,38,39].

White matter feature extraction was performed through the use of standard processing procedures. First, DICOM-formed images were transformed into Nifti-formed images. Subsequently, the dtiInit preprocessing pipeline in the VISTASOFT package (MATLAB toolkit, version 1.0, https://github.com/vistalab/vistasoft) was used to preprocess DTI images for routine preprocessing steps including an eddy current correction, head motion correction, rigid-body alignment to the T1 image, resampling to 2-mm isotropic voxels, skull stripping, and tensor model fitting with a simple least-squares fit method. Second, we performed an AFQ pipeline (version 1.2, https://github.com/yeatmanlab/AFQ) to extract diffusion measurements (FA, MD, RD, and AxD) and morphometric features (fiber core linearity values, curvature, torsion, and volume) at multiple locations (100 points) along the trajectory of 18 major tracts, including the bilateral corticospinal, inferior fronto-occipital fasciculus (IFOF), inferior longitudinal fasciculus (ILF), superior longitudinal fasciculus (SLF), arcuate, cingulum cingulate, uncinate, thalamic radiation, and minor and major callosum forcep tracts. Specifically, there were three steps in the AFQ pipeline: fiber tract identification, cleaning, and quantification. 1), whole-brain fibers were tracked by using the deterministic streamlines tracking algorithm, after which fiber tracts were segmented based on the waypoint ROI procedure and refined via fiber tract probability maps. 2), the abnormal fibers that were longer or deviated far from the core of the fiber tract were iteratively cleaned. 3), each fiber was resampled to 100 equally spaced nodes between the two ROIs, and diffusion properties were calculated at each node of each fiber [33].

However, due to complicated factors such as heterogeneity and image quality, the AFQ method could not guarantee that all 18 fiber bun-

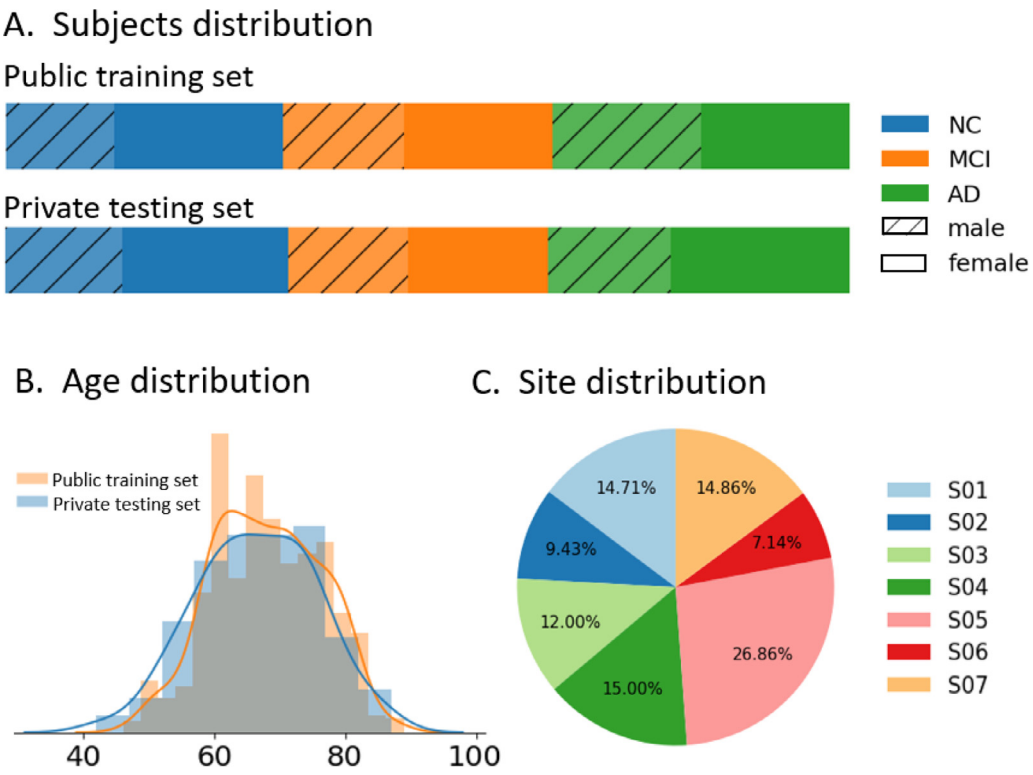

Fig. 1. Demographic statistics (diagnosis class, gender, age, and sites) distributions. (A) Diagnosis and gender ratio on the training set (N=700) and the private testing set (N=125). (B) Age distribution on the two datasets. (C). Site distribution on the two datasets. All of the statistical distributions were matched in the public training set and the private testing set; therefore, only the site distribution of the public dataset is shown in (C).

dles could be tracked for every subject, and we had to balance feature quality (fiber recognition rate) and the number of subjects. Our results showed that the associated number of subjects was 502, 703, 779, 807, and 825 when the threshold at the fiber number was 18, 17, 16, 15, and 14, respectively. Hence, only the subjects with more than 14 identified fiber bundles were retained, which ensured both a high fiber recognition rate (14/18=78%) and the large size of the dataset (N=825, consisting of 276 NCs, 294 ADs, and 255 MCIs). The WM feature dimension of each subject was 14,400 (18 fibers × 100 points × 8 metrics).

2.2. Classification competition

The competition aimed to evaluate and develop an analysis framework for the optimal performance of AD binary classification (AD vs NC) by using diffusion measurements along 18 major white matter tracts that were extracted via AFQ. We also encouraged the challengers to explore triclassification (AD, MCI, and NC).

We launched public training data and invited scientists to submit solutions to predict AD. Briefly, to ensure the matching category and site proportion in the public training set and in the private testing set, we listed all of the data according to the sequence of the diagnosis type within sites. Subsequently, a uniform sampling method was used for each site, 125 individuals were selected for the private testing set, and the remaining samples were selected for the public training set. The public training dataset (containing WM features, age, gender, diagnosis class, and site tag) was provided to the participants (available on a website), and the private testing dataset was unavailable to the participants (the challengers will never be able to obtain the diagnosis label) to blindly evaluate all of the models at the end of the competition. The distribution of the demographic statistics and sites in the public training set and in the private testing set is shown in Fig. 1.

Each team had 48 hours to submit up to 5 different solutions for the final evaluation, and the top ten solutions were awarded money prizes. All of the solutions were ranked according to the order of accuracy, F1 score, and area under the receiver operating characteristic curve (ROC) curve (AUC) [40,41].

3. Results

The competition was conducted for three weeks and attracted 77 registered teams that originated from more than 40 universities/institutes in China, the United States, and the United Kingdom. In total, 48 teams completed the challenge, and 130 solutions were submitted for evaluation.

3.1. The performance of AD prediction with white matter features

Overall, as shown in Fig. 2A, more than 50% of the solutions performed well in the AD and NC binary classifications, achieving a median accuracy of over 74.12%, a F1 score of over 0.73, and a AUC of over 0.83, although some solutions failed the prediction (with accuracies of under 50%). Table 1 lists the performances of the top ten solutions that exhibited excellent discrimination jobs, which obtained an average accuracy of 80.47% (F1 score=0.80, sensitivity=80.91%, and specificity=80.00%), and an average AUC of 0.87 (ranging from 0.81 to 0.89). The best solution achieved a prediction accuracy of 82.35% (with a sensitivity of 86.36% and a specificity of 78.05%). The ROC curves of the top ten models are shown in Fig. 2C.

In addition, 58 solutions of triclassification (AD, MCI, and NC) from 27 teams were submitted. The discrimination ability achieved anaccuracy of 46.46% (ranging from 21.60% to 55.20%), a macroaverage F1-score of 0.44, and a macroaverage AUC of 0.64 (ranging from 0.41 to 0.74). The best solution obtained an accuracy of 55.20%, a macroaverage F1-score of 0.54, and a macroaverage AUC of 0.73. This result is comparable with previous triclassification studies [17,42].

Fig. 2. Performances of the submitted models. The performances of all of the solutions on the locked private test set, including accuracy, F1 score, AUC, sensitivity, and specificity, are shown in (A). The pie chart in (B) displays the usage ratios of the different types of models, among which SVM was most frequently selected. The scattergram in (B) plots the accuracy of the models on the locked private testing set, thus showing that XGBoost, SVM, and logical regression performed better. (C) ROC curves for the top ten models. (D) The accuracy distribution on the public training set and locked private testing set are shown in this figure. The models marked with red dots obtained far higher accuracies on the training set than on the testing set, thus illustrating a severe overfitting problem. The top ten models marked with black dots predicted well on both the training set and testing set.
Abbreviations for (A): ACC – accuracy, SEN – sensitivity, SPE – specificity. Abbreviations for (B): XGB – XGBoost, LR – logistic regression, M-C – multiclassifier integration, DL – deep learning, Per – perceptron, RF – random forest, O – others. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 1
Performance of the top ten models.

Rank	Accuracy	F1 score	Sensitivity	Specificity	AUC
1	82.35%	0.81	86.36%	78.05%	0.88
2	81.18%	0.82	75.00%	87.80%	0.88
3	81.18%	0.81	79.55%	82.93%	0.87
4	81.18%	0.80	81.82%	80.49%	0.88
5	81.18%	0.80	84.09%	78.05%	0.86
6	80.00%	0.79	81.82%	78.05%	0.87
7	80.00%	0.78	84.09%	75.61%	0.81
8	80.00%	0.78	86.36%	73.17%	0.89
9	78.82%	0.80	72.73%	85.37%	0.88
10	78.82%	0.79	77.27%	80.49%	0.85
Average	80.47%	0.80	80.91%	80.00%	0.87

3.2. The generalization ability of the models

From a general view, the performance on the public training dataset (average accuracy=85.30%) was significantly higher than that on the private testing set (average accuracy=70.28%), but both were fairly predictable, as shown in Fig. 2D. As displayed by the red points, some methods performed well on the public training dataset (with an accuracy of over 95%) but poorly on the private testing set (accuracy of less than 60%), thus indicating a serious overfitting problem. Some models performed stably on both the public training dataset and the private testing set (with an accuracy of approximately 80%), such as the top ten models (marked with black dots in Fig. 2D), which demonstrated credible generalization abilities.

3.3. Experience from the competing solutions

A total of 130 solutions were received during the competition. According to the model descriptions that were provided by the participants, 86 (66%) of the solutions used feature reduction and feature selection, and 40 (30%) of the solutions used the ensemble learning method.

When concerning feature engineering, the commonly used feature selection and reduction algorithms included the F-score, principal component analysis, max-relevance and min-redundancy method, and statistical methods (t-test or Pearson correlation). A few teams employed some advanced feature representations that were derived from encoding the original features. In terms of the classification models, the widely used classifiers included SVM (32.31%), logistic regression (11.54%), and XGBoost (7.69%) (Fig. 2B), which could achieve relatively good performance when combined with feature dimensionality reduction. Deep learning models, such as CNN and autoencoder (a total of 15.38%), were also selected but did not perform as excellent as was expected, with the best accuracy of only 77.65%. Some teams adopted ensemble learning to integrate the results of multiple classifiers, but it did not demonstrate obvious advantages, even though only 2 models achieved a ranking in the top ten.

The top 10 solutions were not special, except in regards to the combination of feature engineering and machine learning models, including SVM, logistic regression, and XGBoost. These were also conducted in other solutions with relatively poor performance, thus indicating that the selection of hyperparameters is possibly one of the most significant techniques that is affected by the division of the training set and the

validation set, as well as by the optimization method and selection bias [43].

4. Discussion

Based on one of the largest multisite DTI biobanks, a set of classification methods were collected by the competition. The classification performance between AD and NC of the obtained WM properties via the AFQ method achieved the highest accuracy of 82.35% on the private testing set, thus demonstrating that DTI is a powerful tool for the early detection of AD.

Our results for the challenge are robust with high generalizability because of the evaluation strategy on the private testing set that is both unavailable to the competitor and independent of the training process, thus avoiding uncertain factors that can arise due to overfitting [44], circular analysis [45], or the degrees of freedom of researchers [46,47]. A small sample size is insufficient for independent replication and reanalysis, and intercenter circular pooling cross-validation cannot prevent overfitting on the reused training sets [44,48]. More importantly, the adjustment of the training strategy (according to the performance of the pooled-out testing subjects) can frequently also lead to overfitting [44]. In both cases, the classifiers may be unable to generalize to additional data sets, and the results are often difficult to be replicated [31,49,50]. In contrast, blind evaluations can provide a relatively objective and credible performance, as well as select highly generalized models [31,51]. We believe that the present challenge has a good generalization power, and we theorize that a larger training set and additional independent cohorts may be able to boost classification, as well as the generalization performance.

The bias between the performance of the submitted models in the competition reinforces the fact that we must pay more attention to model selection and parameter adjustment. For overfitting models (the red points displayed in Fig. 2D) that performed well on the public training dataset while performed poorly on the private testing set, the possible reason is that the models were so complex that they captured residual variation or noise as important features, thus decreasing the overall generalizability [52]. Several teams with lower accuracies adopted similar (or even the same) classifiers as the top 10 solutions, but they did not work well, which partially indicates the importance of tuning parameters, which need to be treated with caution to avoid both underfitting and overfitting through appropriate cross-validation and evaluation strategies [44,48]. Deep learning models have been widely used in the computer-aided diagnosis of AD, due to their ability to learn suitable features and robustness [41,53–55]. The reasons for the unexpected mediocre performance of deep learning methods can be complicated. One possible reason may be that the challengers did not use it correctly. Another reason is that, although deep learning models tend to fit the large sample size data well, this do not indicate that they can perform and generalize well in a small sample size. However, there is no need to doubt the effectiveness of deep learning methods in neuroscience because they have been verified by numerous studies [56–59]. Thus, we aim to recruit more labs to join us in searching for more independent cohorts and more solutions for the early detection of AD.

4.1. Limitations, caveats, and future directions

We found conclusive evidence that WM fiber measures based on DTI can be used in the early diagnosis of AD. However, there are still some limitations and additional considerations that need to be overcome and achieved, such as post hoc analyses, data quantity and quality enhancements, as well as better algorithm explorations, which are of great significance for understanding white matter as being a biomarker in AD.

First, we only presented the main results of the competition without adequate interpretation, and a post hoc analysis for model explanation and pathological significance is not yet finished. Given the end-to-end machine learning design, these classifiers often appear to be black boxes that are difficult to interpret for disease severity-associated clinical situations. A post hoc detailed analysis based on the top models is one possible way to understand model results and disease-related pathological mechanisms, including the relationship between model output and the cognitive ability score, as well as the degradation patterns of fiber bundles in AD.

Second, the current dataset is far from sufficient, and a larger multimodality multisite dataset is under preparation. The identification of pathological heterogeneity in AD is challenging, and previous results remain inconclusive [60,61]. Hence, the dataset has to be further expanded for the competition to more thoroughly introduce MRI features into the early diagnosis of AD. Furthermore, we need more samples from more sites with high-quality images to enhance the power of cross-validation to improve the generalization performance of the models, which is of great significance to the progression of the results from pure laboratory research to clinical application. In addition, diverse types of radiological features are required to more sufficiently describe the characteristics of AD. For DTI, more WM features, such as FA, MD of the whole brain, and morphological features of fiber bundles, can be used to more comprehensively express the characteristics of white matter [16,21,62,63]. Additionally, the use of multimodality data that combines structural and functional features is a mainstream trend to facilitate not only a better understanding of pathology but also the precise identification of early AD [38,64].

Finally, better algorithms are strongly expected. Our data are characterized by typically high dimensions and low sample sizes, which could not be effectively overcome via traditional feature selection and reduction, thus resulting in the need for more specific algorithms that can adapt to this type of characteristic or that can directly deal with the original features with good performance [65,66]. In addition, it is still inevitable and difficult to deal with the heterogeneity in data from multiple sites that arise from complicated factors, which causes the inappropriateness of directly mixing data for analysis [67–69]. In this competition, some participants entered the site tags into the classifiers as a feature, but this strategy is not applicable for unknown sites. There are many harmonization methods for DTI images [50,70,71], and it has been proven that the adoption of harmonization methods can partially reduce the sites' effects while also maintaining biological content [69,72,73], which may be beneficial for AD classification and may be worth further explorations, based on our dataset.

The identification of the details of the methods and algorithm codes is valuable for reproducibility and expansibility. To benefit community efforts, we hereby make the data and codes of the top-ranked models openly available (https://github.com/YongLiuLab) and expect new models for the accurate diagnosis of AD.

In summary, the ultimate purpose of conducting and improving this competition is to evaluate the generalization performance of classification methods to promote the progression of laboratory research to clinical practice. For the present challenge, we focused on AD diagnosis, but the classification of multiple mental illnesses would be more valuable for clinical applications in the future, thus providing benefits for both the revelation of pathology differences between diseases and for improving the accuracy of a precise diagnosis. Due to the challenges of large multiple disease datasets and complicated classification models, this large research area needs to be further explored. For example, Sabuncu and colleagues (2015) performed a simple binary classification (patients versus healthy people) based on six sMRI datasets, including AD, schizophrenia, autism, attention deficit, and hyperactivity disorder [74]. Some studies have explored the classification performance between similar diseases, such as schizophrenia, bipolar disorder, and borderline personality disorder [75,76]. To date, some mature multi-disease datasets, such as ENIGMA, and single-disease datasets (such as ADNI, ABIDE, and OASIS, among others) have been released, which can be conveniently and rapidly used for future relevant multiple mental illness studies.

5. Conclusion

In conclusion, we successfully built a multisite DTI image platform and verified the feasibility of white matter measurements for AD diagnosis through a competition. The data set and portions of the codes are available as open sources. The project we present in this study would benefit from having more researchers involved for sharing data or machine learning models, as well as for framing biomarker extraction as an open, international challenge to predict AD with the largest available DTI dataset biobank.

Declaration of Competing Interest

The authors declare no competing financial interests.

Funding

This work was partially supported by the Beijing Natural Science Funds for Distinguished Young Scholars (No. JQ200036), and the National Natural Science Foundation of China (grant nos. 81871438, 81901101), and the Foundation Strengthening Programme (No. 2019-JCJQ-JJ-151), and the Medical Big Data R & D project of PLA general hospital (No. 2018MBD-028).

Data collection and sharing for this project were funded by the National Natural Science Foundation of China (Grant Nos. 61633018, 81571062, 81400890, 81471120, 81701781).

Acknowledgments

The authors express appreciation to Prof. Dong Ming and Prof. Liang Wan from Tianjin University for their kind help with the competition. We gratefully acknowledge the financial support of Shenzhen Hanix United, Ltd. for awards to the winners.

References

[1] L Jia, M Quan, Y Fu, T Zhao, Y Li, C Wei, et al., Dementia in China: epidemiology, clinical management, and research advances, Lancet Neurol. 19 (1) (2020) 81–92, doi:10.1016/S1474-4422(19)30290-X.

[2] J Jia, C Wei, S Chen, F Li, Y Tang, W Qin, et al., The cost of Alzheimer's disease in China and re-estimation of costs worldwide, Alzheimers Dement. 14 (4) (2018) 483–491, doi:10.1016/j.jalz.2017.12.006.

[3] BP. Leifer, Early diagnosis of Alzheimer's disease: clinical and economic benefits, J. Am. Geriatr. Soc. 51 (5 Suppl Dementia) (2003) S281–S288, doi:10.1046/j.1532-5415.5153.x.

[4] A Burns, S. Iliffe, Alzheimer's disease, BMJ 338 (feb05 1) (2009) b158, doi:10.1136/bmj.b158.

[5] JM Long, DM. Holtzman, Alzheimer disease: an update on pathobiology and treatment strategies, Cell 179 (2) (2019) 312–339, doi:10.1016/j.cell.2019.09.001.

[6] S Rathore, M Habes, MA Iftikhar, A Shacklett, C. Davatzikos, A review on neuroimaging-based classification studies and associated feature extraction methods for Alzheimer's disease and its prodromal stages, Neuroimage 155 (2017) 530–548, doi:10.1016/j.neuroimage.2017.03.057.

[7] GB Frisoni, NC Fox, CR Jack Jr., P Scheltens, PM Thompson, The clinical use of structural MRI in Alzheimer disease, Nat. Rev. Neurol. 6 (2) (2010) 67–77, doi:10.1038/nrneurol.2009.215.

[8] A Khazaee, A Ebrahimzadeh, A Babajani-Feremi, Alzheimer's disease neuroimaging I. Classification of patients with MCI and AD from healthy controls using directed graph measures of resting-state fMRI, Behav. Brain Res. 322 (Pt B) (2017) 339–350, doi:10.1016/j.bbr.2016.06.043.

[9] W Li, XF Lin, X. Chen, Detecting Alzheimer's disease based on 4D fMRI: an exploration under deep learning framework, Neurocomputing 388 (2020) 280–287, doi:10.1016/j.neucom.2020.01.053.

[10] Y Ding, JH Sohn, MG Kawczynski, H Trivedi, R Harnish, NW Jenkins, et al., A deep learning model to predict a diagnosis of alzheimer disease by using (18)F-FDG PET of the brain, Radiology 290 (2) (2019) 456–464, doi:10.1148/radiol.2018180958.

[11] Ou YN, Xu W, Li JQ, Guo Y, Cui M, Chen KL, et al. FDG-PET as an independent biomarker for Alzheimer's biological diagnosis: a longitudinal study. Alzheimer's research & therapy 2019; 11(1): 57; https://doi.org/ 10.1186/s13195-019-0512-1.

[12] EE Bron, M Smits, WM van der Flier, H Vrenken, F Barkhof, P Scheltens, et al., Standardized evaluation of algorithms for computer-aided diagnosis of dementia based on structural MRI: the CADDementia challenge, Neuroimage 111 (2015) 562–579, doi:10.1016/j.neuroimage.2015.01.048.

[13] RV Marinescu, NP Oxtoby, AL Young, EE Bron, AW Toga, MW Weiner, et al., TADPOLE challenge: accurate Alzheimer's disease prediction through crowdsourced forecasting of future data, Predict. Intell. Med. 11843 (2019) 1–10, doi:10.1007/978-3-030-32281-6_1.

[14] GI Allen, N Amoroso, C Anghel, V Balagurusamy, CJ Bare, D Beaton, et al., Crowdsourced estimation of cognitive decline and resilience in Alzheimer's disease, Alzheimers Dement. 12 (6) (2016) 645–653, doi:10.1016/j.jalz.2016.02.006.

[15] A Sarica, A Cerasa, A Quattrone, V. Calhoun, Editorial on special issue: machine learning on MCI, J. Neurosci. Methods 302 (2018) 1–2, doi:10.1016/j.jneumeth.2018.03.011.

[16] Y Jin, C Huang, M Daianu, L Zhan, EL Dennis, RI Reid, et al., 3D tract-specific local and global analysis of white matter integrity in Alzheimer's disease, Hum. Brain Mapp. 38 (3) (2017) 1191–1207, doi:10.1002/hbm.23448.

[17] X Dou, H Yao, F Feng, P Wang, B Zhou, D Jin, et al., Characterizing white matter connectivity in Alzheimer's disease and mild cognitive impairment: an automated fiber quantification analysis with two independent datasets, Cortex 129 (2020) 390–405, doi:10.1016/j.cortex.2020.03.032.

[18] Y Zhang, N Schuff, AT Du, HJ Rosen, JH Kramer, ML Gorno-Tempini, et al., White matter damage in frontotemporal dementia and Alzheimer's disease measured by diffusion MRI, Brain 132 (Pt 9) (2009) 2579–2592, doi:10.1093/brain/awp071.

[19] SE Rose, AL Janke, JB. Chalk, Gray and white matter changes in Alzheimer's disease: a diffusion tensor imaging study, J. Magn. Reson. Imag. 27 (1) (2008) 20–26, doi:10.1002/jmri.21231.

[20] D Medina, L DeToledo-Morrell, F Urresta, JD Gabrieli, M Moseley, D Fleischman, et al., White matter changes in mild cognitive impairment and AD: a diffusion tensor imaging study, Neurobiol. Aging 27 (5) (2006) 663–672, doi:10.1016/j.neurobiolaging.2005.03.026.

[21] E Horgusluoglu-Moloch, G Xiao, M Wang, Q Wang, X Zhou, K Nho, et al., Systems modeling of white matter microstructural abnormalities in Alzheimer's disease, NeuroImage Clin. 26 (2020) 102203, doi:10.1016/j.nicl.2020.102203.

[22] M Dyrba, F Barkhof, A Fellgiebel, M Filippi, L Hausner, K Hauenstein, et al., Predicting prodromal Alzheimer's disease in subjects with mild cognitive impairment using machine learning classification of multimodal multicenter diffusion-tensor and magnetic resonance imaging data, J. Neuroimaging 25 (5) (2015) 738–747, doi:10.1111/jon.12214.

[23] M Dyrba, M Ewers, M Wegrzyn, I Kilimann, C Plant, A Oswald, et al., Robust automated detection of microstructural white matter degeneration in Alzheimer's disease using machine learning classification of multicenter DTI data, PLoS One 8 (5) (2013) e64925, doi:10.1371/journal.pone.0064925.

[24] SJ Teipel, M Wegrzyn, T Meindl, G Frisoni, AL Bokde, A Fellgiebel, et al., Anatomical MRI and DTI in the diagnosis of Alzheimer's disease: a European multicenter study, J. Alzheimers Dis. 31 (Suppl 3) (2012) S33–S47, doi:10.3233/JAD-2012-112118.

[25] G Prasad, SH Joshi, TM Nir, AW Toga, PM Thompson, Alzheimer's disease neuroimaging I. Brain connectivity and novel network measures for Alzheimer's disease classification, Neurobiol. Aging. 36 (Suppl 1) (2015) S121–S131, doi:10.1016/j.neurobiolaging.2014.04.037.

[26] CY Wee, PT Yap, W Li, K Denny, JN Browndyke, GG Potter, et al., Enriched white matter connectivity networks for accurate identification of MCI patients, Neuroimage 54 (3) (2011) 1812–1822, doi:10.1016/j.neuroimage.2010.10.026.

[27] H Chen, X Sheng, R Qin, C Luo, M Li, R Liu, et al., Aberrant white matter microstructure as a potential diagnostic marker in Alzheimer's disease by automated fiber quantification, Front. Neurosci. 14 (956) (2020), doi:10.3389/fnins.2020.570123.

[28] A Ebadi, JL Dalboni da Rocha, DB Nagaraju, F Tovar-Moll, I Bramati, G Coutinho, et al., Ensemble classification of Alzheimer's disease and mild cognitive impairment based on complex graph measures from diffusion tensor images, Front. Neurosci. 11 (2017) 56, doi:10.3389/fnins.2017.00056.

[29] F Wang, LP Casalino, D. Khullar, Deep learning in medicine-promise, progress, and challenges, JAMA Intern. Med. 179 (3) (2019) 293–294, doi:10.1001/jamainternmed.2018.7117.

[30] BA Richards, TP Lillicrap, P Beaudoin, Y Bengio, R Bogacz, A Christensen, et al., A deep learning framework for neuroscience, Nat. Neurosci. 22 (11) (2019) 1761–1770, doi:10.1038/s41593-019-0520-2.

[31] CW Woo, LJ Chang, MA Lindquist, TD. Wager, Building better biomarkers: brain models in translational neuroimaging, Nat. Neurosci. 20 (3) (2017) 365–377, doi:10.1038/nn.4478.

[32] JD Yeatman, A Richie-Halford, JK Smith, A Keshavan, A. Rokem, A browser-based tool for visualization and analysis of diffusion MRI data, Nat. Commun. 9 (1) (2018) 940, doi:10.1038/s41467-018-03297-7.

[33] JD Yeatman, RF Dougherty, NJ Myall, BA Wandell, HM. Feldman, Tract profiles of white matter properties: automating fiber-tract quantification, PLoS One 7 (11) (2012) e49790, doi:10.1371/journal.pone.0049790.

[34] S Teubner-Rhodes, KI Vaden Jr., SL Cute, JD Yeatman, RF Dougherty, MA Eckert, Aging-resilient associations between the arcuate fasciculus and vocabulary knowledge: microstructure or morphology? J. Neurosci. 36 (27) (2016) 7210–7222, doi:10.1523/JNEUROSCI.4342-15.2016.

[35] X Zhang, Y Sun, W Li, B Liu, W Wu, H Zhao, et al., Characterization of white matter changes along fibers by automated fiber quantification in the early stages of Alzheimer's disease, NeuroImage Clin. 22 (2019) 101723, doi:10.1016/j.nicl.2019.101723.

[36] H Sun, S Lui, L Yao, W Deng, Y Xiao, W Zhang, et al., Two patterns of white matter abnormalities in medication-naive patients with first-episode schizophrenia revealed by diffusion tensor imaging and cluster analysis, JAMA Psychiatry 72 (7) (2015) 678–686, doi:10.1001/jamapsychiatry.2015.0505.

[37] MD Sacchet, G Prasad, LC Foland-Ross, SH Joshi, JP Hamilton, PM Thompson, et al., Structural abnormality of the corticospinal tract in major depressive disorder, Biol. Mood Anxiety Disord. 4 (1) (2014) 8, doi:10.1186/2045-5380-4-8.

[38] D Jin, P Wang, A Zalesky, B Liu, C Song, D Wang, et al., Grab-AD: generalizability and reproducibility of altered brain activity and diagnostic classification in Alzheimer's Disease, Hum. Brain Mapp. 41 (12) (2020) 3379–3391, doi:10.1002/hbm.25023.

[39] J Li, D Jin, A Li, B Liu, C Song, P Wang, et al., ASAF: altered spontaneous activity fingerprinting in Alzheimer's disease based on multisite fMRI, Sci. Bull. 64 (14) (2019) 998–1010, doi:10.1016/j.scib.2019.04.034.
[40] K Zhao, YH Ding, Y Han, Y Fan, AF Alexander-Bloch, T Han, et al., Independent and reproducible hippocampal radiomic biomarkers for multisite Alzheimer's disease: diagnosis, longitudinal progress and biological basis, Sci. Bull. 65 (13) (2020) 1103–1113, doi:10.1016/j.scib.2020.04.003.
[41] D Jin, B Zhou, Y Han, J Ren, T Han, B Liu, et al., Generalizable, reproducible, and neuroscientifically interpretable imaging biomarkers for Alzheimer's disease, Adv. Sci. 7 (14) (2020) 2000675, doi:10.1002/advs.202000675.
[42] R Cuingnet, E Gerardin, J Tessieras, G Auzias, S Lehéricy, M-O Habert, et al., Automatic classification of patients with Alzheimer's disease from structural MRI: a comparison of ten methods using the ADNI database, Neuroimage 56 (2) (2011) 766–781, doi:10.1016/j.neuroimage.2010.06.013.
[43] AF Mendelson, MA Zuluaga, M Lorenzi, BF Hutton, S Ourselin, Alzheimer's disease neuroimaging I. Selection bias in the reported performances of AD classification pipelines, NeuroImage Clin. 14 (2017) 400–416, doi:10.1016/j.nicl.2016.12.018.
[44] M Hosseini, M Powell, J Collins, C Callahan-Flintoft, W Jones, H Bowman, et al., I tried a bunch of things: the dangers of unexpected overfitting in classification of brain data, Neurosci. Biobehav. Rev. 119 (2020) 456–467, doi:10.1016/j.neubiorev.2020.09.036.
[45] N Kriegeskorte, WK Simmons, PS Bellgowan, CI. Baker, Circular analysis in systems neuroscience: the dangers of double dipping, Nat. Neurosci. 12 (5) (2009) 535–540, doi:10.1038/nn.2303.
[46] JP. Ioannidis, Why most published research findings are false, PLoS Med. 2 (8) (2005) e124, doi:10.1371/journal.pmed.0020124.
[47] KS Button, JP Ioannidis, C Mokrysz, BA Nosek, J Flint, ES Robinson, et al., Power failure: why small sample size undermines the reliability of neuroscience, Nat. Rev. Neurosci. 14 (5) (2013) 365–376, doi:10.1038/nrn3475.
[48] G. Varoquaux, Cross-validation failure: small sample sizes lead to large error bars, Neuroimage 180 (Pt A) (2018) 68–77, doi:10.1016/j.neuroimage.2017.06.061.
[49] SA Iqbal, JD Wallach, MJ Khoury, SD Schully, JP. Ioannidis, Reproducible research practices and transparency across the biomedical literature, PLoS Biol. 14 (1) (2016) e1002333, doi:10.1371/journal.pbio.1002333.
[50] M Yu, K Linn, P Cook, M Phillips, M McInnis, M Fava, et al., Statistical harmonization corrects site effects in functional connectivity measurements from multi-site fMRI data, Hum. Brain Mapp. 39 (11) (2018) 4213–4227, doi:10.1002/hbm.24241.
[51] RA Poldrack, G Huckins, G. Varoquaux, Establishment of best practices for evidence for prediction: a review, JAMA Psychiatry 77 (5) (2020) 534–540, doi:10.1001/jamapsychiatry.2019.3671.
[52] S Mutasa, S Sun, R. Ha, Understanding artificial intelligence based radiology studies: what is overfitting? Clin. Imaging 65 (2020) 96–99, doi:10.1016/j.clinimag.2020.04.025.
[53] C Lian, M Liu, J Zhang, D. Shen, Hierarchical fully convolutional network for joint atrophy localization and Alzheimer's disease diagnosis using structural MRI, IEEE Trans. Pattern Anal. Mach. Intell. 42 (4) (2020) 880–893, doi:10.1109/TPAMI.2018.2889096.
[54] D Shen, G Wu, HI. Suk, Deep learning in medical image analysis, Annu. Rev. Biomed. Eng. 19 (2017) 221–248, doi:10.1146/annurev-bioeng-071516-044442.
[55] S Qiu, PS Joshi, MI Miller, C Xue, X Zhou, C Karjadi, et al., Development and validation of an interpretable deep learning framework for Alzheimer's disease classification, Brain 143 (6) (2020) 1920–1933, doi:10.1093/brain/awaa137.
[56] J Sui, M Liu, JH Lee, J Zhang, V. Calhoun, Deep learning methods and applications in neuroimaging, J. Neurosci. Methods 339 (2020) 108718, doi:10.1016/j.jneumeth.2020.108718.
[57] MA Schulz, BTT Yeo, JT Vogelstein, J Mourao-Miranada, JN Kather, K Kording, et al., Different scaling of linear models and deep learning in UKBiobank brain images versus machine-learning datasets, Nat. Commun. 11 (1) (2020) 4238, doi:10.1038/s41467-020-18037-z.
[58] G Martensson, D Ferreira, T Granberg, L Cavallin, K Oppedal, A Padovani, et al., The reliability of a deep learning model in clinical out-of-distribution MRI data: a multicohort study, Med. Image Anal. 66 (2020) 101714, doi:10.1016/j.media.2020.101714.
[59] Y Liu, A Jain, C Eng, DH Way, K Lee, P Bui, et al., A deep learning system for differential diagnosis of skin diseases, Nat. Med. 26 (6) (2020) 900–908, doi:10.1038/s41591-020-0842-3.
[60] M Habes, MJ Grothe, B Tunc, C McMillan, DA Wolk, C. Davatzikos, Disentangling heterogeneity in Alzheimer's disease and related dementias using data-driven methods, Biol. Psychiatry 88 (1) (2020) 70–82, doi:10.1016/j.biopsych.2020.01.016.
[61] M Ten Kate, E Dicks, PJ Visser, WM van der Flier, CE Teunissen, F Barkhof, et al., Atrophy subtypes in prodromal Alzheimer's disease are associated with cognitive decline, BrainBrain 141 (12) (2018) 3443–3456, doi:10.1093/brain/awy264.
[62] R Mito, T Dhollander, Y Xia, D Raffelt, O Salvado, L Churilov, et al., In vivo microstructural heterogeneity of white matter lesions in healthy elderly and Alzheimer's disease participants using tissue compositional analysis of diffusion MRI data, NeuroImage Clin. 28 (2020) 102479, doi:10.1016/j.nicl.2020.102479.
[63] R Raja, G Rosenberg, A. Caprihan, Review of diffusion MRI studies in chronic white matter diseases, Neurosci. Lett. 694 (2019) 198–207, doi:10.1016/j.neulet.2018.12.007.
[64] M De Marco, L Beltrachini, A Biancardi, AF Frangi, A. Venneri, Machine-learning support to individual diagnosis of mild cognitive impairment using multimodal MRI and cognitive assessments, Alzheimer Dis. Assoc. Disord. 31 (4) (2017) 278–286, doi:10.1097/WAD.0000000000000208.
[65] CF Tsai, YT. Sung, Ensemble feature selection in high dimension, low sample size datasets: parallel and serial combination approaches, Knowl.-Based Syst. 203 (2020) 106097, doi:10.1016/j.knosys.2020.106097.
[66] SJ Raudys, AK. Jain, Small sample-size effects in statistical pattern-recognition - recommendations for practitioners, IEEE Trans. Pattern Anal. Mach. Intell. 13 (3) (1991) 252–264, doi:10.1109/34.75512.
[67] H Ni, V Kavcic, T Zhu, S Ekholm, J. Zhong, Effects of number of diffusion gradient directions on derived diffusion tensor imaging indices in human brain, AJNR Am. J. Neuroradiol. 27 (8) (2006) 1776–1781, doi:10.1080/02841850600816331.
[68] C Vollmar, J O'Muircheartaigh, GJ Barker, MR Symms, P Thompson, V Kumari, et al., Identical, but not the same: intra-site and inter-site reproducibility of fractional anisotropy measures on two 3.0T scanners, Neuroimage 51 (4) (2010) 1384–1394, doi:10.1016/j.neuroimage.2010.03.046.
[69] C Wachinger, A Rieckmann, S Polsterl, Alzheimer's disease neuroimaging I, the Australian imaging B, lifestyle flagship study of a. Detect and correct bias in multi-site neuroimaging datasets, Med. Image Anal. 67 (2021) 101879, doi:10.1016/j.media.2020.101879.
[70] JP Fortin, D Parker, B Tunc, T Watanabe, MA Elliott, K Ruparel, et al., Harmonization of multi-site diffusion tensor imaging data, Neuroimage 161 (2017) 149–170, doi:10.1016/j.neuroimage.2017.08.047.
[71] MS Pinto, R Paolella, T Billiet, P Van Dyck, PJ Guns, B Jeurissen, et al., Harmonization of brain diffusion MRI: concepts and methods, Front. Neurosci. 14 (2020) 396, doi:10.3389/fnins.2020.00396.
[72] C. Davatzikos, Machine learning in neuroimaging: progress and challenges, Neuroimage 197 (2019) 652–656, doi:10.1016/j.neuroimage.2018.10.003.
[73] JP Fortin, N Cullen, YI Sheline, WD Taylor, I Aselcioglu, PA Cook, et al., Harmonization of cortical thickness measurements across scanners and sites, Neuroimage 167 (2018) 104–120, doi:10.1016/j.neuroimage.2017.11.024.
[74] MR Sabuncu, E Konukoglu, Alzheimer's disease neuroimaging I. Clinical prediction from structural brain MRI scans: a large-scale empirical study, Neuroinformatics 13 (1) (2015) 31–46, doi:10.1007/s12021-014-9238-1.
[75] HG Schnack, M Nieuwenhuis, NEM van Haren, L Abramovic, TW Scheewe, RM Brouwer, et al., Can structural MRI aid in clinical classification? A machine learning study in two independent samples of patients with schizophrenia, bipolar disorder and healthy subjects, Neuroimage 84 (2014) 299–306, doi:10.1016/j.neuroimage.2013.08.053.
[76] I Perez Arribas, GM Goodwin, JR Geddes, T Lyons, KEA. Saunders, A signature-based machine learning model for distinguishing bipolar disorder and borderline personality disorder, Transl. Psychiatry 8 (1) (2018) 274-https://doi.org/, doi:10.1038/s41398-018-0334-0.