Journal of Alzheimer's Disease xx (2022) x–xx
DOI 10.3233/JAD-215654
IOS Press



# A Stacking Framework for Multi-Classification of Alzheimer's Disease Using Neuroimaging and Clinical Features

Durong Chen, Fuliang Yi, Yao Qin, Jiajia Zhang, Xiaoyan Ge, Hongjuan Han, Jing Cui, Wenlin Bai, Yan Wu and Hongmei Yu* the Alzheimer's Disease Neuroimaging Initiative[1]
*Department of Health Statistics, School of Public Health, Shanxi Medical University, Taiyuan, China*

Accepted 4 April 2022
Pre-press 27 April 2022

**Abstract**.
**Background:** Alzheimer's disease (AD) is a severe health problem. Challenges still remain in early diagnosis.
**Objective:** The objective of this study was to build a Stacking framework for multi-classification of AD by a combination of neuroimaging and clinical features to improve the performance.
**Methods:** The data we used were from the Alzheimer's Disease Neuroimaging Initiative database with a total of 493 subjects, including 125 normal control (NC), 121 early mild cognitive impairment, 109 late mild cognitive impairment (LMCI), and 138 AD. We selected structural magnetic resonance imaging (sMRI) features by voting strategy. The imaging features, demographic information, Mini-Mental State Examination, and Alzheimer's Disease Assessment Scale-Cognitive Subscale were combined together as classification features. We proposed a two-layer Stacking ensemble framework to classify four types of people. The first layer represented support vector machine, random forests, adaptive boosting, and gradient boosting decision tree; the second layer was a logistic regression classifier. Additionally, we analyzed performance of only sMRI feature and combined features and compared the proposed model with four base classifiers.
**Results:** The Stacking model combined with sMRI and non-imaging features outshined four base classifiers with an average accuracy of 86.96%. Compared with using sMRI data alone, sMRI combined with non-imaging features significantly improved diagnostic accuracy, especially in NC versus LMCI and LMCI versus AD by 14.08%.
**Conclusion:** The Stacking framework we used can improve performance in diagnosis of AD using combined features.

Keywords: Alzheimer's disease, classification, ensemble learning, neuroimaging



## INTRODUCTION

Alzheimer's disease (AD) is a progressive neurodegenerative disease with cognitive decline and physical impairment [1], and millions worldwide continue to suffer from AD [2]. The development of effective treatments remains stalled, under certain situation medical field emphasizes early diagnosis [3]. Mild cognitive impairment (MCI) is a transitional stage between the normal controls (NC) and AD [4]. Individuals with MCI develop to AD with a conversion at an annual rate of 5–25% [5]. In order to define an earlier onset of disease, MCI can be divided into early mild cognitive impairment (EMCI)

[1]Data used in this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (http://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/wpcontent/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf.

*Correspondence to: Hongmei Yu, Department of Health Statistics, School of Public Health, Shanxi Provincial Key Laboratory of Major Diseases Risk Assessment, Shanxi Medical University, 56 South XinJian Road, Taiyuan, 030001, P.R. China. Tel.: +86 351 4135049; Fax: +86 351 2027943; E-mail: yu@sxmu.edu.cn.

and late mild cognitive impairment (LMCI) according to Alzheimer's Disease Neuroimaging Initiative (ADNI) database [6]. Some studies have proven that individuals with LMCI are more subjected to convert to AD than individuals with EMCI [7]. MCI versus AD classification itself is a more difficult problem than distinguishing between AD and NC, because MCI diagnosis is stuck at a gray area and can be easily confused with AD or NC [8]. Recently, more research criteria have been proposed for early diagnosis of AD or MCI [9], which plays a vital role in timely prevention and treatment of AD.

Current diagnosis standards depend on neuropsychological assessments and brain imaging techniques for individuals with AD. Neuropsychological assessments are simple and practical, especially in the elderly at the community or in areas with poor medical conditions. Among brain imaging techniques, structure magnetic resonance imaging (sMRI) is a safe, non-invasive, and objective technology, which produces high resolution spatial images. Based on cerebral atrophy, sMRI can offer reliable information about the progression of AD [10]. Region-based analysis methods are employed to examine volume of brain and detect shrinkage of brain tissue for detecting differences in images [11]. The key of region-based analysis methods is the determination of the region of interest (ROI) which can be utilized to identify the anatomical differences to assist diagnosis [12].

Though tremendous progress has been made in diagnosis of AD, an imprecise diagnostic environment still exists. On the one hand, most of the existing studies choose only one classifier or compare several classifiers and screen the best one as the final classifier. Classifiers boast their own advantages and call for some specific applications. Gray et al. classified AD, MCI, and NC only by random forest (RF) [13]. Ezzati et al. applied six machine learning methods: decision trees (DT), support vector machines (SVM), K-nearest neighbor, ensemble linear discriminant, boosted trees, and RF to classify NC and AD; the best model was used for predicting clinical outcome of MCI [14]. Zhe et al. also only selected adaptive boosting (AdaBoost) to complete the classification task [15]. On the other hand, multi-classification still faces lower accuracy. Jin et al. used DT to classify NC, MCI, and AD with an accuracy of 56.52% [16]. Son et al. classified AD, MCI, and NC using RF and MRI, and the accuracy was 53.33% [17]. Zhe et al. adapted AdaBoost to distinguish AD, MCI, and NC with an accuracy of 75.76% [15]. The existing studies about multi-classification have the poor diagnostic performance, which may result in diagnostic errors in clinical settings.

It could be helpful to combine several classifiers to enhance diagnostic performance for classification of AD. Stacking, an ensemble method, combines different base classifiers into one meta-classifier, which proves simplicity and high performance with combined capability of different classifiers [18, 19]. In this study, we designed a Stacking framework to build a multi-classification (NC/EMCI/LMCI/AD) by combining sMRI, neuropsychological assessments, and demographic information as to enhance performance of the diagnosis. In the first layer, four base classifiers included SVM, RF, AdaBoost, and gradient boosting decision tree (GBDT). We chose logistic regression (LR) in the second layer to fuse outputs of first layer and get the final result of classification.

## METHODS

### *ADNI dataset*

The data we used were from the ADNI database (http://adni.loni.usc.edu). The ADNI, a public-private partnership, was launched in 2003 by Michael W. Weiner, and subjects were recruited from USA and Canada. The primary goal of ADNI is to test whether the serial MRI, PET, other biological markers, and neuropsychological assessments can be combined to measure the progression of MCI and early AD. The identification of sensitive and specific markers of early AD progression is intended to aid researchers and clinicians to develop new treatments and monitor their effectiveness, as well as lessening time and cost of clinical trials. The research protocol was approved by each the local ethical committee and the written informed consent was obtained from each participant. For more information, see http://www.adni-info.org.

### *Participants*

There are four stages of ADNI, and the data we used were mainly derived from ADNI-2&GO. There were 493 participants in our study, 125 NC, 121 EMCI, 109 LMCI, and 138 AD, respectively, whose baseline MRIs were available. Demographic information comprised age, gender (male/female), years of education, and marital status (married/single (unmarried, divorced, widowed)). The neuropsychological assessments we used included the Mini-Mental State Examination (MMSE) and

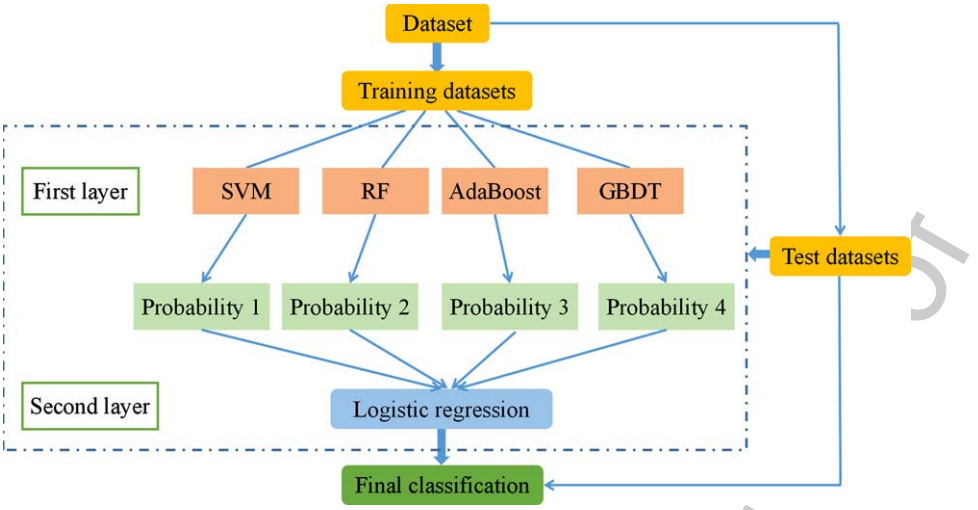


Fig. 1. The Stacking framework design.

the Alzheimer's Disease Assessment Scale (ADAS-Cog11). These non-imaging features, known as risk factors of AD, can be easily obtained by non-AD specialists. We defined the above two part of non-imaging features as clinical information.

*Image data preprocessing*

All subjects were scanned by the 3.0T MR scan and the parameters were defined TR = 2300 ms, TE = 2.98 ms, flip angle = 90°, thickness = 1.2 mm, [FOV] = 240 × 240 $mm^2$ and matrix size = 256 × 256. We used Statistical Parametric Mapping (SPM12) on MATLAB platform for preprocessing. The original sMRI images were converted from DICOM to NIFTI format. We used the Montreal Neurological Institute space for spatial normalization. sMRI images were segmented into three different tissues. Our work focused on gray matter. The sMRI was divided into 90 brain regions using the automatic anatomical labeling (AAL) template in REST software, and gray matter volume (GMV) was extracted. The corresponding names of AAL brain template subdivisions are shown in Supplementary Table 1, where an odd number indicates the left brain and an even number indicates the right brain.

*Feature selection*

In the neuroimaging community, reduction of features is a critical and essential process before training the model. The main purpose of this process is to select the most relevant features and remove redundant ones to avoid over-fitting in models. Feature selection methods are divided into the following three ones: filter, wrapper, and embedded [20]. However, previous studies have argued that wrapper and embedded are superior to filter methods in neuroimaging data [21, 22]. In this study, we employed support vector machine recursive feature elimination (SVM-RFE), LR based on L1 regularization, and GBDT to select image features. The final features were selected more than once to ensure better stability and less redundancy.

*Classification models*

We applied six binary classifications to achieve multi-classification. The six binary classification tasks were NC versus EMCI, NC versus LMCI, NC versus AD, EMCI versus LMCI, EMCI versus AD, and LMCI versus AD.

The purpose of ensemble is to combine multiple algorithms to improve performance. Figure 1 illustrates the structure of Stacking procedure. Stacking contains two layers. In the first layer, the individual classification models represented heterogeneity are trained on training sets. The base classifiers in the first layer take two requirements. The first one involves high diversity and the second one emphasizes high accuracy. In this study, we used four base classifiers: SVM, RF, AdaBoost, and GBDT in the first layer, which have different modeling ideas and good performance in cross-validation. The four parallel results of classifiers were calculated. In the second layer, the meta learner should have strong generalization

Table 1
Demographics of participants

| Characteristics | NC | EMCI | LMCI | AD | $\chi^2/F$ | $p$ |
|---|---|---|---|---|---|---|
| N | 125 | 121 | 109 | 138 | | |
| Age (mean ± SD)/y | 73.53 ± 6.37 | 70.71 ± 6.74 | 71.45 ± 7.03 | 74.73 ± 8.19 | 8.49 | <0.001 |
| Gender (n (%)) | | | | | | |
| Male | 65 (48.00%) | 71 (58.68%) | 59 (54.13%) | 80 (57.97%) | 3.67 | 0.302 |
| Female | 60 (52.00%) | 50 (41.32%) | 50 (45.87%) | 58 (42.03%) | | |
| Education (mean ± SD)/y | 16.56 ± 2.54 | 16.00 ± 2.59 | 16.52 ± 2.58 | 15.67 ± 2.68 | 3.56 | 0.014 |
| Marriage status (n (%)) | | | | | | |
| Married | 81 (64.80%) | 94 (77.69%) | 83 (76.15%) | 119 (86.23%) | 16.88 | <0.001 |
| Single | 44 (35.20%) | 27 (22.31%) | 26 (23.85%) | 19 (13.77%) | | |
| MMSE (mean ± SD) | 29.03 ± 1.24 | 28.43 ± 1.54 | 27.64 ± 1.79 | 23.09 ± 2.13 | 328.87 | <0.001 |
| ADAS-Cog11 (mean ± SD) | 5.94 ± 3.10 | 7.46 ± 3.17 | 11.70 ± 3.17 | 20.93 ± 7.18 | 246.33 | <0.001 |

MMSE, Mini-Mental State Examination; ADAS-Cog11, Alzheimer's Disease Assessment Scale contains 11 items; SD, standard deviation.

ability to correct the bias of base learners and avoid over-fitting [23, 24]. Hence, the LR was trained in the second layer to fuse classifying outputs from the first layer. Finally, the classification results were calculated using the test sets, and evaluation indicators were established for classification performance. And we used RF to rank the importance of selected brain regions and clinical features in each classification.

*Performance metrics*

We used the Scikit-Learn machine learning library in Python 3.8.5 software to build AD classification diagnostic models. Nested cross-validation was applied in this study. To evaluate the performance of classifiers, the four indicators: accuracy (ACC), recall, F1 score and the area under the ROC curve (AUC) were chosen. DeLong's test was used to verify AUC. The parameters are listed in Supplementary Table 2.

## RESULTS

*Clinical information*

Table 1 summarizes participants' clinical characteristics. Among them, 275 (55.78%) were male and 218 (44.22%) were female; the mean age was 72.71 ± 7.31 years, with range from 55 to 90 years old. The mean years of education was 16.16 ± 2.60 years, ranging from 9 to 20 years. As for the marital status, 377 (76.47%) were married, 116 (23.53%) were single (unmarried, divorced, widowed). The mean score of MMSE stood at 26.92 ± 2.97 and ADAS-Cog11 represented 11.78 ± 7.81. Beyond gender, these features were statistical significance, which were selected into the AD classification model.

*Selected features*

The specific results of features extracted by three feature selection methods at each binary classification task are summarized in Supplementary Table 3. The brain regions retained at each binary classification task after feature selections by voting strategy are shown in Table 2. The main brain regions that eventually entered classification models were hippocampus, parahippocampal gyrus, amygdala, superior limbic gyrus, thalamus, middle temporal gyrus, and inferior temporal gyrus. Furthermore, the correlations between these screened brain regions and clinical information showed that the majority of brain regions were associated with age, ADAS-Cog11, and MMSE, see Supplementary Tables 4–9 for specific results.

*Classification results*

Our Stacking framework could perform multi-classification of AD using sMRI data and a combination of sMRI and non-imaging features. NC versus AD data were cited as an example to explain the classification results. The output metrics of four base classifiers and Stacking are detailed in Table 3. When sMRI was taken as classification feature only, the result of four base classifiers presented relatively ordinary performance, and ACC, recall, AUC and F1 score were below 88%. However, compared with base classifiers, Stacking boasted a better performance especially in AUC. The similar results were showed in sMRI joined together with non-imaging features. The performance metrics of four base classifiers were all below 98%, while Stacking outperformed base classifiers as it was represented by 0.9999 of AUC particularly.

Compared with only sMRI, the Stacking of sMRI joined together with non-imaging features in NC

Table 2
The results of feature voting

| Stage | Image features |
|---|---|
| NC versus EMCI | Frontal_Sup_L, Frontal_Sup_Medial_R, Rectus_L, Cingulum_Mid_R, Cingulum_Post_L, Hippocampus_R, Parahippocampal_L, Parahippocampal_R, Calcarine_L, Cuneus_L, Lingual_L, Lingual_R, Occipital_Inf_R, SupraMarginal_L, Paracentral_Lobule_R, Pallidum_L, Thalamus_R, Heschl_R, Temporal_Pole_Mid_L, Temporal_Pole_Mid_R, Temporal_Inf_L |
| NC versus LMCI | Hippocampus_L, Hippocampus_R, Parahippocampal_R, Amygdala_R, Calcarine_L, Lingual_L, Lingual_R, Occipital_Mid_L, Postcentral_R, Putamen_L, Thalamus_L, Temporal_Inf_R |
| NC versus AD | Hippocampus_L, Hippocampus_R, Parahippocampal_R, Amygdala_L, Amygdala_R, Calcarine_L, Lingual_L, SupraMarginal_L, Angular_L, Thalamus_L, Temporal_Mid_L, Temporal_Inf_L |
| EMCI versus LMCI | Frontal_Sup_R, Olfactory_L, Rectus_R, Cingulum_Mid_R, Cingulum_Post_L, Hippocampus_L, Hippocampus_R, Amygdala_L, Amygdala_R, Cuneus_R, Occipital_Mid_L, SupraMarginal_R, Precuneus_L, Thalamus_L, Temporal_Sup_R, Temporal_Mid_L |
| EMCI versus AD | Cingulum_Mid_R, Cingulum_Post_L, Hippocampus_L, Hippocampus_R, Parahippocampal_R, Amygdala_L, Amygdala_R, Occipital_Sup_L, Occipital_Mid_R, Occipital_Inf_R, Precuneus_L, Thalamus_L, Thalamus_R, Temporal_Mid_L, Temporal_Inf_L |
| LMCI versus AD | Supp_Motor_Area_L, Frontal_Sup_Medial_L, Hippocampus_L, Parahippocampal_L, Amygdala_L, Amygdala_R, Fusiform_L, Postcentral_R, Parietal_Inf_L, SupraMarginal_L, SupraMarginal_R, Temporal_Mid_L, Temporal_Inf_L |

Table 3
The classification result of NC versus AD

| | Classifier | ACC | recall | AUC | F1 score |
|---|---|---|---|---|---|
| sMRI | SVM | 0.8666 | 0.8675 | 0.8629 | 0.8707 |
| | RF | 0.8515 | 0.8573 | 0.8466 | 0.8550 |
| | AdaBoost | 0.8171 | 0.8047 | 0.8126 | 0.8215 |
| | GBDT | 0.8554 | 0.8496 | 0.8503 | 0.8568 |
| | **Stacking** | **0.8937** | **0.8768** | **0.9522** | **0.8966** |
| sMRI+clinical information | SVM | 0.9734 | 0.9788 | 0.9748 | 0.9736 |
| | RF | 0.9771 | 0.9631 | 0.9766 | 0.9779 |
| | AdaBoost | 0.9657 | 0.9791 | 0.9622 | 0.9704 |
| | GBDT | 0.9733 | 0.9767 | 0.9699 | 0.9751 |
| | **Stacking** | **0.9873** | **0.9836** | **0.9999** | **0.9895** |

versus AD were increased to different degrees, where recall and ACC value increased by 10.68% and 9.36%. The sMRI combined with non-imaging features showed significant advantages, especially in the NC versus LMCI stage with an increase in recall of up to 19.19%. Figure 2 shows the results of Stacking using sMRI data alone and sMRI combined with non-imaging features in each classification. To verify the efficacy of sMRI joined together with non-imaging features, we performed Delong's test for AUC values in the Stacking results, which confirmed statistical significance except NC versus EMCI. Of note, the classification result of features combined with sMRI and non-imaging features outperformed only sMRI feature.

The results of the other five binary classifications showed the same effect, as described in Supplementary Tables 10–14. And the result of first ten importance features are shown in Fig. 3. The results indicated that ADAS-Cog11 plays an important role in all classifications.

## DISCUSSION

In our study, we designed a Stacking framework to improve performance of multi-classification based on the sMRI and combined features, and the result of features with sMRI and clinical information exhibited better classification ability than only sMRI feature.

In general, there is inherent conflict between accuracy and diversity of individual learners, the more diversity, the less accuracy. Actually, it is hoped that different base learners can be "accurate but different". In this study, four base learners were selected as the first layer learner. Among them, SVM is a single classifier, while RF, AdaBoost, and GBDT are three different ensemble models. Comparison of these five models, the experimental results showed that our Stacking framework can achieve strong complementarity between different base learners. The most pronounced result was EMCI versus LMCI classification using combined features, which had the most significant increase in AUC value. However, in

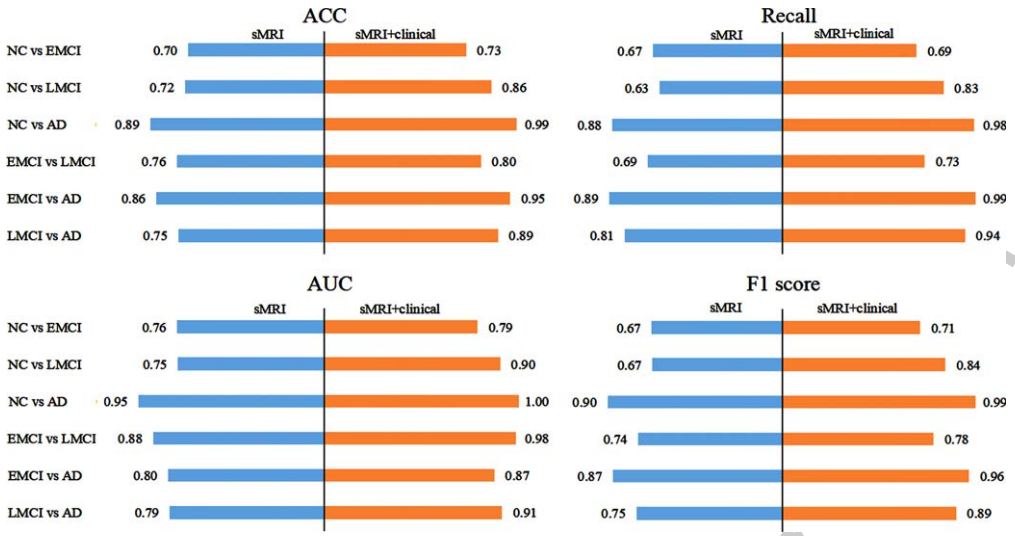


Fig. 2. The results of Stacking classification.

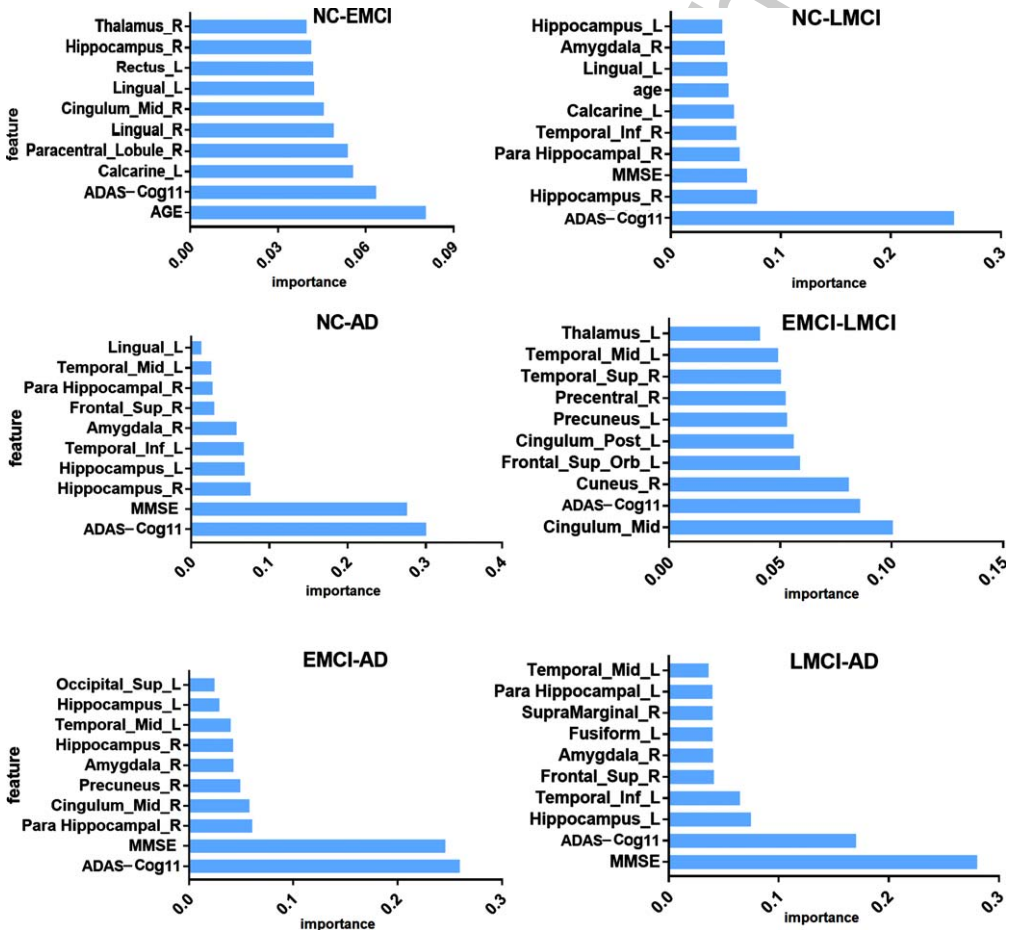


Fig. 3. The feature importance ranking.

Table 4
Classification performance of existing studies

| Study | Subjects | Algorithm | Features | Overall ACC |
|---|---|---|---|---|
| Ebadi et al. [26] | 15AD/15MCI/15CT | Ensemble learning by voting strategy | Diffusion Tensor Imaging | 80.00% (AD versus CT)<br>83.30% (AD versus MCI)<br>70.00% (MCI versus CT) |
| Sorensen et al. [27] | 100AD/100MCI/ 100cMCI/100NC | Ensemble SVM using linear kernel | sMRI, age, sex, and MMSE | 55.6% |
| | | Ensemble SVM using radial basis function (RBF) kernel | | 55.0% |
| Gray et al. [13] | 37AD/34sMCI/ 41pMCI/35HC | RF | MRI<br>PDG-PET<br>CSF | 89.00% (AD versus HC)<br>74.60% (MCI versus HC)<br>58.40% (sMCI versus pMCI) |
| Our framework | 138AD/109LMCI/ 121EMCI/125NC | Stacking | sMRI, neuropsychological assessments and demographic information | 72.70% (NC versus EMCI)<br>85.63% (NC versus LMCI)<br>98.73% (NC versus AD)<br>80.29% (EMCI versus LMCI)<br>95.38% (EMCI versus AD)<br>89.01% (LMCI versus AD) |

CT, healthy subjects; cMCI, converting MCI; sMCI, MCI individuals who have progressed to AD; pMCI, MCI individuals who have so far remained stable; PDG-PET, positron emission tomography imaging with the radiotracer [$^{18}$F]-fluorodeoxyglucose; CSF, cerebrospinal fluid.

NC versus AD and LMCI versus AD, the Stacking classification results of AUC using combined features provided only a modest boost compared with base classifiers. This is owing to the progression of AD, the brain structure gaps are most evident, which lead to a higher performance of four base classifiers. We integrated the four base classifiers that had achieved high classification performances thus the performance did not improve significantly.

Our study adapted multiple binary classifications for the purpose of multi-classification. Regardless of whether Stacking framework were used, the classification of NC versus AD had the best effect using combined features, and AUC values reached more than 96% in the base classifiers, which can be understood that atrophy of brain structures does differ in the NC and AD. Compared with previous related studies, our work performed with higher accuracy, as shown in Table 4. In our study, each binary classification with combined features showed good discriminative ability and overall ACC of classification was 86.96%.

Our study also found that using combined features can produce more powerful classifiers compared to using sMRI feature alone. Especially in NC versus LMCI classification, the recall of Stacking was increased by 19.19%. Most likely, the cognitive status and clinical information of these two stages are quite different. This prompts us to take early cognitive intervention for LMCI subjects. In EMCI versus LMCI, the performance of these two statuses had more subtle improvement using sMRI combined with non-imaging features and they were subjected to classification difficulties. The possible reason is that EMCI and LMCI belong to the MCI status, the differences from degree of atrophy in brain and clinical cognition between the two are small making it more difficult to distinguish. Motter et al. also found that there were no statistically significant differences between the EMCI and LMCI groups in terms of lesion volume [25]. In our results, the growth of AUC in NC versus EMCI have no statistical difference in combined features, which likely due to the fact that the total contribution of clinical features was smaller than sMRI. This signals the need for physicians to spend more effort in differentiating between these two statuses. Taken together, combined features could provide more information about the likelihood of cognitive impairment.

The sMRI scans with high analytical accuracy show changes in brain structures monitoring AD process [28], since the typical distribution of gray matter atrophy revealed by sMRI may achieve better diagnostic accuracy [28, 29]. As such, in this study, we mainly focused on GMV changes of sMRI in AD analysis. Our research found the main brain atrophies were located in hippocampus, parahippocampal gyrus, amygdala, and temporal lobe. The behavioral studies have shown that learning and memory storage and retrieval play a critical role in the hippocampus [30, 31]. The parahippocampal gyrus is also associated with memory storage and retrieval. Echavarri used sMRI to distinguish NC, aMCI, and AD finding that the difference in parahippocampal volume atrophy was greater than that in hippocampus [32]. The amygdala is associated with emotion, learning, and memory involving in the processing of

long-term memory and consolidating memory storage in other parts of the brain [33]. In addition to the brain regions mentioned above, the thalamus, superior limbic gyrus, middle temporal lobes, and inferior temporal lobes were also associated with disease progression by feature selection. The thalamus is the higher center of sensation, and it is also bound up with memory function and emotion regulation. The thalamus is involved in many different neuronal pathways and its function is closely related to motor behavior, emotion, motivation, association, and cognitive ability [34]. The temporal lobe is primarily relevant to hearing, language comprehension, memory, and mental activity. Atrophy of brain tissue is a long process and occurs in the hippocampus and internal olfactory cortex firstly, and then affects the parietal, temporal, and frontal lobes. It has been documented that temporal lobe atrophy is exacerbated in individuals with MCI and AD [25, 35]. Besides, there is an association between these brain regions and clinical features (age, ADAS-Cog11, and MMSE) and such an association might be considered as a new biomarker or might provide evidence to be considered for future studies.

The goal of our study is to ultimately create an automated machine learning and find biomarkers to help physicians to make more streamlined and accurate diagnoses. The Stacking framework we designed has significant translational potential in AD, which can help physicians by offering an objective assessment and a second opinion. In addition, our framework can be applied to other diseases, such as Parkinson's disease. The combination of features from different modalities may considerably increase the potential of AD diagnosis. These medical examinations can be easily obtained and used for early screening of AD in the community. This will not only reduce the burden on society and families, but also promote early detection of AD achieving a reasonable allocation of social resources. In such cases, our model may aid non-invasive monitoring of AD development. Furthermore, the development of methods which efficiently combines multimodal features is a field to be explored by next studies.

However, some limitations also remain in our study. First, this study extracted GMV from sMRI as morphological characteristics, demographic information, and neuropsychological assessments. Further studies will focus on incorporating multiple features such as fMRI, DTI, PET, CSF, and genes. Second, other base classifiers, ensemble learning algorithms, and construction strategies could be incorporated to analyze cognitive decline in the elderly and provide new references to assist clinical diagnosis. Third, we reported accuracy of Stacking framework in training and testing datasets, which showed slight over-fitting, and specific results are available in Supplementary Table 15. In addition, we will use external validation to generalize the stability of performance. These limitations should be addressed in our future studies.

In conclusion, we used six binary classification tasks to achieve the purpose of multi-classification. The Stacking framework combining SVM, RF, AdaBoost, and GBDT model was employed to classify NC/EMCI/LMCI/AD based on sMRI and non-imaging features. The performance of our Stacking framework was improved significantly, and the result of combined features outperformed only sMRI feature. The model we constructed in this study provides an approach for the future translation of neuroimaging into AD benefit.

## ACKNOWLEDGMENTS

This work is supported by the National Natural Science Foundation of China (Grant number 81973154) and the Natural Science Foundation for Young Scientists of Shanxi Province (Grant number 202103021223242).

Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE HealtNCare; 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 (http://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. Before interviewing each participant, written informed consent including aims and methods such as physical and neurological examinations were obtained from all participants. The authors are also grateful to the participants for their support and cooperation in making this research possible.

Authors' disclosures available online (https://www.j-alz.com/manuscript-disclosures/21-5654r1).

## SUPPLEMENTARY MATERIAL

The supplementary material is available in the electronic version of this article: https://dx.doi.org/10.3233/JAD-215654.

## REFERENCES

[1] DeTure MA, Dickson DW (2019) The neuropathological diagnosis of Alzheimer's disease. *Mol Neurodegener* **14**, 32.

[2] (2020) 2020 Alzheimer's disease facts and figures. *Alzheimers Dement* **16**, 391-460.

[3] (2019) The need for early detection and treatment in Alzheimer's disease. *EBioMedicine* **9**, 1-2.

[4] Anderson ND (2019) State of the science on mild cognitive impairment (MCI). *CNS Spectr* **24**, 78-87.

[5] Pennanen C, Testa C, Laakso MP, Hallikainen M, Helkala EL, Hanninen T, Kivipelto M, Kononen M, Nissinen A, Tervo S, Vanhanen M, Vanninen R, Frisoni GB, Soininen H (2005) A voxel based morphometry study on mild cognitive impairment. *J Neurol Neurosurg Psychiatry* **76**, 11-14.

[6] Jessen F, Wolfsgruber S, Wiese B, Bickel H, Mosch E, Kaduszkiewicz H, Pentzek M, Riedel-Heller SG, Luck T, Fuchs A, Weyerer S, Werle J, van den Bussche H, Scherer M, Maier W, Wagner M, German Study on Aging, Cognition and Dementia in Primary Care Patients (2014) AD dementia risk in late MCI, in early MCI, and in subjective memory impairment. *Alzheimers Dement* **10**, 76-83.

[7] Zhang T, Zhao Z, Zhang C, Zhang J, Jin Z, Li L (2019) Classification of early and late mild cognitive impairment using functional brain network of resting-state fMRI. *Front Psychiatry* **10**, 572.

[8] Mateos-Perez JM, Dadar M, Lacalle-Aurioles M, Iturria-Medina Y, Zeighami Y, Evans AC (2018) Structural neuroimaging as clinical predictor: A review of machine learning applications. *Neuroimage Clin* **20**, 506-522.

[9] Dubois B, Feldman HH, Jacova C, DeKosky ST, Barberger-Gateau P, Cummings J, Delacourte A, Galasko D, Gauthier S, Jicha G, Meguro K, O'Brien J, Pasquier F, Robert P, Rossor M, Salloway S, Stern Y, Visser PJ, Scheltens P (2007) Research criteria for the diagnosis of Alzheimer's disease: Revising the NINCDS–ADRDA criteria. *Lancet Neurol* **6**, 734-746.

[10] Aisen PS, Petersen RC, Donohue M, Weiner MW; Alzheimer's Disease Neuroimaging Initiative (2015) Alzheimer's Disease Neuroimaging Initiative 2 Clinical Core: Progress and plans. *Alzheimers Dement* **11**, 734-739.

[11] Egger K, Rau A, Yang S, Klppel S, Urbach HJ (2020) Automated voxel- and region-based analysis of gray matter and cerebrospinal fluid space in primary dementia disorders. *Brain Res* **1739**, 146800.

[12] Liu J, Wang J, Tang Z, Hu B, Wu FX, Pan Y (2018) Improving Alzheimer's disease classification by combining multiple measures. *IEEE/ACM Trans Comput Biol Bioinform* **15**, 1649-1659.

[13] Gray KR, Aljabar P, Heckemann RA, Hammers A, Rueckert D, Alzheimer's Disease Neuroimaging Initiative (2013) Random forest-based similarity measures for multi-modal classification of Alzheimer's disease. *Neuroimage* **65**, 167-175.

[14] Ezzati A, Zammit AR, Harvey DJ, Habeck C, Hall CB, Lipton RB, Alzheimer's Disease Neuroimaging Initiative (2019) Optimizing machine learning methods to improve predictive models of Alzheimer's disease. *J Alzheimers Dis* **71**, 1027-1036.

[15] Wang Z, Zheng Y, Zhu DC, Bozoki AC, Li T (2018) Classification of Alzheimer's disease, mild cognitive impairment and normal control subjects using resting-state fMRI based network connectivity analysis. *IEEE J Transl Eng Health Med* **6**, 1801009.

[16] Jin M, Deng W (2018) Predication of different stages of Alzheimer's disease using neighborhood component analysis and ensemble decision tree. *J Neurosci Methods* **302**, 35-41.

[17] Son SJ, Kim J, Park H (2017) Structural and functional connectional fingerprints in mild cognitive impairment and Alzheimer's disease patients. *PLoS One* **12**, e0173426.

[18] Menahem E, Rokach L, Elovici Y (2009) Troika - An improved stacking schema for classification tasks. *Inform Sci* **179**, 4097-4122.

[19] Naimi AI, Balzer LB (2018) Stacked generalization: An introduction to super learning. *Eur J Epidemiol* **33**, 459-464.

[20] Mwangi B, Tian TS, Soares JC (2014) A review of feature reduction techniques in neuroimaging. *Neuroinformatics* **12**, 229-244.

[21] Tangaro S, Amoroso N, Brescia M, Cavuoti S, Chincarini A, Errico R, Inglese P, Longo G, Maglietta R, Tateo A, Riccio G, Bellotti R (2015) Feature selection based on machine learning in MRIs for hippocampal segmentation. *Comput Math Methods Med* **2015**, 814104.

[22] Tohka J, Moradi E, Huttunen H, Alzheimer's Disease Neuroimaging Initiative (2016) Comparison of feature selection techniques in machine learning for anatomical brain MRI in dementia. *Neuroinformatics* **14**, 279-296.

[23] Wang Y, Wang D, Geng N, Wang Y, Yin Y, Jin Y (2019) Stacking-based ensemble learning of decision trees for interpretable prostate cancer detection. *Appl Soft Comput* **77**, 188-204.

[24] Tang J, Liang J, Han C, Li Z, Huang H (2019) Crash injury severity analysis using a two-layer Stacking framework. *Accid Anal Prev* **122**, 226-238.

[25] Motter JN, Pelton GH, D'Antonio K, Rushia SN, Pimontel MA, Petrella JR, Garcon E, Ciovacco MW, Sneed JR,

Doraiswamy PM, Devanand DP (2018) Clinical and radiological characteristics of early versus late mild cognitive impairment in patients with comorbid depressive disorder. *Int J Geriatr Psychiatry* **33**, 1604-1612.

[26] Ebadi A, Dalboni da Rocha JL, Nagaraju DB, Tovar-Moll F, Bramati I, Coutinho G, Sitaram R, Rashidi P (2017) Ensemble classification of Alzheimer's disease and mild cognitive impairment based on complex graph measures from diffusion tensor images. *Front Neurosci* **11**, 56.

[27] Sorensen L, Nielsen M, Alzheimer's Disease Neuroimaging Initiative (2018) Ensemble support vector machine classification of dementia using structural MRI and mini-mental state examination. *J Neurosci Methods* **302**, 66-74.

[28] Krajcovicova L, Klobusiakova P, Rektorova I (2019) Gray matter changes in Parkinson's and Alzheimer's disease and relation to cognition. *Curr Neurol Neurosci Rep* **19**, 85.

[29] Albert MS, DeKosky ST, Dickson D, Dubois B, Feldman HH, Fox NC, Gamst A, Holtzman DM, Jagust WJ, Petersen RC, Snyder PJ, Carrillo MC, Thies B, Phelps CH (2011) The diagnosis of mild cognitive impairment due to Alzheimer's disease: Recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimers Dement* **7**, 270-279.

[30] Knierim JJ (2015) The hippocampus. *Curr Biol* **25**, R1116-1121.

[31] Mu Y, Gage FH (2011) Adult hippocampal neurogenesis and its role in Alzheimer's disease. *Mol Neurodegener* **6**, 85.

[32] Echavarri C, Aalten P, Uylings HB, Jacobs HI, Visser PJ, Gronenschild EH, Verhey FR, Burgmans S (2011) Atrophy in the parahippocampal gyrus as an early biomarker of Alzheimer's disease. *Brain Struct Funct* **215**, 265-271.

[33] Poulin SP, Dautoff R, Morris JC, Barrett LF, Dickerson BC, Alzheimer's Disease Neuroimaging Initiative (2011) Amygdala atrophy is prominent in early Alzheimer's disease and relates to symptom severity. *Psychiatry Res* **194**, 7-13.

[34] de Jong LW, van der Hiele K, Veer IM, Houwing JJ, Westendorp RG, Bollen EL, de Bruin PW, Middelkoop HA, van Buchem MA, van der Grond J (2008) Strongly reduced volumes of putamen and thalamus in Alzheimer's disease: An MRI study. *Brain* **131**, 3277-3285.

[35] Fjell AM, McEvoy L, Holland D, Dale AM, Walhovd KB, Alzheimer's Disease Neuroimaging Initiative (2014) What is normal in normal aging? Effects of aging, amyloid and Alzheimer's disease on the cerebral cortex and the hippocampus. *Prog Neurobiol* **117**, 20-40.