Aging as an active player in Alzheimer's Disease Classification: Insights from feature selection in BrainAge Models.

Jorge Garcia Condado[a,b,c], Ines Verdugo Recuero[d], Iñigo Tellaetxe Elorriaga[a,b], Colin Birkhenbil[c], Maria Carrigan[c,e,f,g], Ibai Diez[a,c,j], Rachel F Buckley[c,h,i], Asier Erramuzpe[a,j], Jesus M Cortes[a,j,f,+] for the Alzheimer's Disease Neuroimaging Initiative*

[a] Computational Neuroimaging Lab, Biobizkaia Health Research Institute, Barakaldo, 4890, Spain

[b] Biomedical Research Doctorate Program, University of the Basque Country, Leioa, 48940, Spain

[c] Department of Neurology, Massachusetts General Hospital, Harvard Medical School, Boston, MA, 02114, USA

[d] Department of Signal Processing and Communications, University Carlos III of Madrid, Leganes, 28911, Spain

[e] Alzheimer Center Amsterdam, Neurology, Vrije Universiteit Amsterdam UMC location VUmc, Amsterdam, The Netherlands

[f] Amsterdam Neuroscience, Neurodegeneration, Amsterdam, The Netherlands

[g] Faculty of Science, Swammerdam Institute for Life Sciences, University of Amsterdam, Amsterdam, The Netherlands

[h] Center for Alzheimer's Research and Treatment, Brigham & Women's Hospital, HMS, Boston

[i] Melbourne School of Psychological Sciences, University of Melbourne, Melbourne, VIC, 3010, Australia

[j] IKERBASQUE, the Basque Foundation for Science, Bilbao, 48009, Spain

[f] Department of Cell Biology and Histology, University of the Basque Country, Leioa, 48940, Spain

[+] Corresponding author: *Email address:* jesus.m.cortes@gmail.com (J.M. Cortes)

* 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

Keywords: Brain Age, Alzheimer's Disease, Neuroimaging, Neuropsychology

Abstract

Background

BrainAge models estimate the biological age of the brain using neuroimaging or clinical features, making them promising tools for studying neurodegenerative diseases like Alzheimer's disease (AD). However, the reliance of BrainAge models on neuroimaging features such as gray matter volume and hippocampal atrophy, can introduce biases linked to individuals' ages as these features are influenced both by normal aging and AD progression. The potential presence of such age-biases raises a critical question: can BrainAge models trained to estimate biological brain aging make meaningful contributions to AD diagnosis, or does any introduced age-bias conflate aging effects with disease pathology? Understanding how deliberate feature selection impacts this confounding effect is essential for developing reliable age-related biomarkers.

Methodology

We ranked neuroimaging and neuropsychological features based on their mutual information with age and their discriminative power across clinical groups (cognitively normal, Mild Cognitive Impairment [MCI], Alzheimer's Disease, and stable vs. progressive MCI). Iteratively, we trained BrainAge models using different subsets of these features—some optimized for predicting aging and others for discrimination of clinical AD stages. We assess the error in BrainAge delta (the difference between predicted biological age and chronological age) and evaluate its classification performance across clinical groups. Finally, we compare using deltas for classification with a logistic regression model directly trained on the neuroimaging features used in the BrainAge models.

Results

Neuroimaging features are more strongly correlated with aging, while neuropsychological features exhibit greater discriminative power for AD classification. BrainAge models optimized for age prediction yield deltas that are suboptimal when used for classifying AD, whereas models trained to generate deltas optimized to be used for classifying AD have reduced age prediction accuracy. This trade-off suggests that BrainAge models may not optimally separate aging-related changes from disease-specific alterations. BrainAge models have varying classification accuracy as compared to direct utilization of features in logistic regression. However, BrainAge provides a continuous measure, offering a single output that can be used across clinical stages, in contrast to classification approaches that require explicit labels for each disease stage.

Conclusion

Aging significantly affects BrainAge-based classification of Alzheimer's disease. Feature selection plays a critical role in mitigating this effect, as the outputs of models trained to predict age, the deltas, may fail in AD classification. These findings underscore the need for task-specific feature selection and model design to ensure that BrainAge models are appropriately applied in neurodegenerative disease research.

Highlights

- Neuroimaging features are effective for age prediction, while neuropsychological tests are better at distinguishing clinical groups, reflecting their distinct roles.
- Ordering features by their mutual information with age improves initial Mean Absolute Error (MAE) of BrainAge models, whereas ordering features by mutual information of their discriminative power enhances the use of deltas for classification but worsens the age prediction accuracy of BrainAge models.
- BrainAge deltas encapsulate disease progression but do not consistently improve classification accuracy compared to using the features directly in a classifier, highlighting the need for further research on their utility.
- All methods are integrated into AgeML, an Open-Source tool for age modeling for reproduction and validation purposes and to encourage further research in diverse areas beyond the brain.

1. Introduction

BrainAge models are predictive tools developed using machine learning or deep learning techniques to estimate an individual's chronological age based on neuroimaging or other biological data [1]. BrainAge models are often employed to assess deviations from typical aging trajectories, with the difference between a participant's predicted and actual age, referred to as the BrainAge delta, serving as a potential biomarker for various diseases. These models have gained significant attention in the study of neurodegenerative diseases, such as Alzheimer's disease (AD) [2] as well as psychiatric disorders and other neurological conditions [3].

Despite the potential of BrainAge models, several challenges hinder their generalizability and clinical utility. One critical issue is the inherent bias introduced by training models on sample populations that are not representative of the broader population. For instance, most models are developed using data from predominantly white individuals, which can result in biased predictions when applied to diverse populations [4,5]. Additionally, the definition of "healthy" in age modeling is dependent on specific studies, and this variability in training data can further skew results [6,7].

Another major challenge is the reproducibility of neuroimaging-based models, especially across different imaging sites and protocols [8,9]. Variability in image acquisition, preprocessing, and analysis methods can introduce noise, complicating efforts to develop models that are robust and generalizable [10]. Although initiatives like the Brain Age Standardized Evaluation (BASE) [11] aim to standardize BrainAge prediction workflows, there remains a need for consensus on evaluation criteria and the consistent application of these standards.

Our previous work has demonstrated that BrainAge models with poorer age prediction accuracy can, at times, yield deltas that perform better when used as input in downstream tasks, such as

classifying between stable Mild Cognitive Impairment (sMCI) and progressive MCI (pMCI) to AD [12]. This suggests that optimizing models purely for age prediction may not necessarily lead to better outcomes when using the deltas to distinguish between individuals with AD dementia and cognitively normal individuals. Despite aging being a strong predictor of dementia, it is non-specific to any one disease, making it challenging to distinguish between normal and pathological ageing.

In this work, we aim to explore how the choice of features used in BrainAge models affects both the error in age prediction and the performance of deltas when use as input in auxiliary downstream tasks, such as classifying between clinical groups. Specifically, we investigate these relationships in the context of AD to elucidate the trade-offs between prediction accuracy and classification performance. We rank neuroimaging and neuropsychological features based on their relationship with age and their discriminative power across clinical groups. By constructing BrainAge models with different feature sets, we assess the error in BrainAge delta and its classification performance when used as input to a logistic regressor. Finally, we compare the predictive value of BrainAge delta to a direct classification approach using neuroimaging features.

2. Materials and Methods

In this study, we explore two different approaches to BrainAge modeling, see Fig. 1. Our first approach explores how BrainAge models trained on different feature sets compare in their age prediction error compared to the use of the generated deltas in the downstream task of classifying two different clinical groups. Our second approach explores whether using the BrainAge deltas for classification is better than using the features directly with a logistic regression for classification.

Both methodologies have been integrated for easy use within AgeML [13]. AgeML is an OpenSource Python package for Age Modeling with Machine Learning. The first approach is integrated under the command ***model_feature_influence***, see Fig. 1 Top panel. The second approach is integrated under the command ***age_model_vs_logistic_regression***, see Fig. 1 Bottom panel. The OpenSource code can be found in the GitHub repository: github.com/compneurobilbao/ageml.

2.1 Data

Data used in the preparation of this article was obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu) [14]. All participants in the ADNI2 and ADNI3 phases who had an initial visit with T1-weighted imaging and neuropsychological evaluation were extracted from the ADNI database. This included clinically normal older adults (CN), and MCI and AD participants. The demographics are provided in Table 1.

Using longitudinal data, we identified those who progressed from MCI to AD. The model was tested in the task of distinguishing between sMCI participants and pMCI participants. sMCI participants were considered stable if they continued to be diagnosed with MCI after 3 years from the initial visit. Similarly, only participants who progressed to AD within 3 years from the initial visit were considered pMCI.

Models were built using structural brain features extracted from T1-weighted images and neuropsychological features at baseline. Ten neuroimaging volumetric features were obtained from T1-weighted images from each participant's first visit. Three of the ten features were extracted with the Structural Image Evaluation with Normalisation of Atrophy Cross-sectional (SIENAX) [15], part of FMRIB Software Library (FSL) [16] to obtain gray matter volume, white matter

volume, cerebrospinal fluid volume, as well as a volume scale value. Then the other seven features were extracted using the FMRIB's Integrated Registration and Segmentation Tool (FIRST) [17] to segment and calculate the volumes of the thalamus, caudate, putamen, pallidum, hippocampus, amygdala, and accumbens over both hemispheres. The volume scale value was used to control for differences in brain size. Six neuropsychological features were obtained from neuropsychological assessments from each participant's first visit. This consisted of scores from standard neuropsychological tests: Mini-Mental State Examination (MMSE)[18], Alzheimer's Disease Assessment Scale (ADAS)[19], Functional Assessment Questionaire (FAQ) [20], and Montreal Cognitive Assessment (MoCA) [21], as well as two metrics generated in the ADNI study: ADNI Memory score [22] and ADNI Executive Function [23]. We compare 4 different clinical groups in the study: CN vs MCI, CN vs AD, MCI vs AD and sMCI vs pMCI.

2.2 Age prediction and classification accuracy for different feature sets

To create the different feature sets, we first order the features based on mutual information (MI), which is well known for capturing relationships beyond the linear dependencies detected by pairwise correlations. In particular, defined from information theory, MI quantifies the statistical dependency between variables, providing a measure of how much knowing one variable reduces uncertainty about another. We use MI to rank features according to two criteria. First, features are ordered by their MI with age using the function *mutual_information_regression* from sklearn [24]. This ranking reflects the extent to which each feature individually correlates with age. Second, features are ranked based on their discriminative power between two clinical groups by calculating their MI with clinical labels using the function *mutual_information_classif* from sklearn [24]. This approach identifies features most relevant for distinguishing between clinical labels. For each different case, feature sets are created by starting with the feature with the highest mutual

information and iteratively adding the next feature with the next highest mutual information. In this study, we use 10 neuroimaging features (grey matter, white matter, cerebrospinal fluid, thalamus, caudate, putamen, pallidum, hippocampus, amygdala, and accumbens normalized volumes) and 6 neuropsychological features (MMSE, ADAS, FAQ, MoCA, ADNI Memory and ADNI executive), so a total of 16 feature sets are created.

For each feature set, we train a BrainAge model based on healthy controls. Specific details on the model training can be found in the Methods section of AgeML [13]. In summary, with y as age, $\boldsymbol{X}$ as our feature set and $f()$ as our pipeline, we optimize the parameters of our pipeline $f()$ by means of minimizing the mean squared error of age on cognitively normal controls. Our pipeline consists of a feature scaler and a linear regressor and we use a 5-fold cross-validation scheme. The reported age prediction error, the Mean Absolute Error (MAE), is reported before the age bias correction step [25].

Afterwards we apply the BrainAge model to the two clinical groups of interest to obtain predicted ages. We then calculate age deltas, the predicted age after age bias correction minus the chronological age, for each participant. If one of the clinical groups is the control group, the predicted ages from the cross-validation out-of-fold predictions are used. Then the deltas are used as input into a logistic regressor to classify between two groups and obtain Areas Under the Curve (AUC). To avoid biased AUC metrics resulting from unequal sample sizes across clinical groups, the larger group is under-sampled to match the number of participants in the smaller group. To address potential circularity in using cognitive test to build BrainAge models whose deltas are later used for classification, we trained the BrainAge model exclusively on healthy controls, preventing data leakage from MCI or AD subjects as argued in previous studies [12].

For each BrainAge model trained on each feature set, we obtain a MAE for the age prediction task and an AUC from using the deltas in the downstream classification task. We then plot the MAE and AUC over the number of features used in training for each model and repeat this process for the two different types of ordering.

2.3 Classification accuracy for different models

We also examine how classification accuracy varies across different models. Specifically, we train four logistic regression models using different inputs. Three models use BrainAge deltas as input to a logistic regressor, where the deltas are derived from three different BrainAge models: a linear regressor, a Ridge regressor, and an SVM. In addition, we train a separate logistic regression model directly on the feature sets. For this analysis, only neuroimaging features are used, as some neuropsychological test features may have been considered by clinicians when assigning clinical labels. Excluding these features helps prevent potential biases in classification when using directly the features in classification.

We are also interested in understanding the value of the BrainAge deltas in terms of whether it adds extra information to the classification task. Therefore, we also train 4 distinct logistic regressors and compare their AUC. The 4 logistic regressors use the following as input: all neuroimaging features, all neuroimaging features and the age of participants, only the BrainAge delta, all neuroimaging features and the BrainAge delta.

3. Results

3.1 Age prediction error and classification accuracy for different feature sets

The results of the feature ranking are shown in Table 2, listed in descending order of importance. The “Age” column refers to the relationship with age, while the remaining columns reflect the

discriminatory power of each feature between clinical groups. We look at discrimination between 4 scenarios: CN vs AD, CN vs MCI, MCI vs AD and sMCI vs pMCI. Neuroimaging features were found to be more informative for inferring age, whereas neuropsychological features proved more effective in discriminating between clinical groups. When mapping MI with age and the discriminative power of each feature in Fig. 2, neuroimaging features show higher MI with age than with classification, whereas neuropsychological features show lower MI with age than with classification. However, neuropsychological features also show low MI when attempting to distinguish between sMCI and pMCI because at baseline these two subgroups are both characterized as MCI and hence, share similar cognitive profiles.

After obtaining the different feature rankings, the pipeline is executed four times, once for each of the four classification tasks of interest: CN vs AD, CN vs MCI, MCI vs AD and sMCI vs pMCI. Results are shown in Figure 3. This figure illustrates how MAE and AUC evolve as more features are added to each set, depending on whether the features are more relevant for age modeling or for classifying clinical groups. When features are ranked by their relevance to age, the initial models trained on the top-ranked subset tend to predict age more accurately, but the deltas perform worse in classification tasks. We see the opposite effect when the features are ranked in importance to discriminate. However, when classifying sMCI and pMCI, we observe that the age-based and discrimination-based models begin to overlap earlier as more features are added because neuropsychological features show MI values comparable to those of neuroimaging features in distinguishing between these two groups.

3.2 Logistic Regression using features vs BrainAge deltas

We trained three different BrainAge models: a linear regressor, a ridge regressor and a SVM with different neuroimaging features. We compared the classification performance of BrainAge deltas

with that of using the original features directly in a logistic regressor. The results for all four scenarios --CN vs AD, CN vs MCI, MCI vs AD and sMCI vs pMCI-- are shown in Fig. 4. Using the original features as input to the logistic regressor yielded the best performance.

Finally, using all neuroimaging features, we evaluated whether adding age or the delta as an additional feature could improve classification performance. The results for each scenario are summarized in Table 3. Adding neither age nor the delta appeared to improve the AUC, as most values remained within one standard deviation of each other. Repeating the experiment with different random seeds yielded fluctuating results, indicated limited impact of these additional features.

4. Discussion

This study aimed to evaluate how different neuroimaging and neuropsychological features influence BrainAge predictions and the utility of BrainAge deltas in classifying Alzheimer's disease (AD) participants. By analyzing the relationship between these features and both age estimation and clinical group distinctions, we assessed the reliability of BrainAge deltas as biomarkers. Our findings provide insight into the strengths and limitations of using BrainAge models for AD classification, highlighting key factors that impact their performance and the need for careful feature selection based on specific research and clinical objectives.

As expected, ranking features by their age prediction accuracy and discriminative power highlights the distinct contributions of neuroimaging and neuropsychological features. Neuroimaging features, specifically gray matter and particularly those related to brain structures like the hippocampus, thalamus, and amygdala, are highly effective in predicting age. In contrast, neuropsychological scores excel at discriminating between clinical groups. This supports the idea

that age prediction relies more on the structural integrity of brain regions, while cognitive performance metrics are more sensitive to the pathological distinctions between clinical groups. It is important to consider possible circularity in using BrainAge deltas built with neuropsychological testing features for classification due to their role in clinical diagnosis, i.e. assigning healthy control, MCI, and AD labels in the first place. In ADNI for example, diagnostic assessment of AD staging includes a selection of these tests in combination with cutoff points and other clinical assessments [14]. In the present study, this concern was accounted for by training our BrainAge model on healthy controls and, and thus preventing data leakage in terms of biases in the BrainAge deltas, as the model is not trained on MCI or AD. In the specific case of classifying sMCI and pMCI, despite neuropsychological tests contributing to initial diagnostic classification, our approach avoids circularity by use of baseline scores to predict future conversion rather than reclassifying subjects based on later assessments [12]. These findings indicate that neuropsychological tests carry predictive information beyond their traditional role in diagnostic labeling based on cutoffs, supporting their validity for assessing and monitoring disease progression.

Notably, the hippocampus is consistently identified for its high discriminative power across all clinical groups. The hippocampus is among the earliest regions affected by AD pathology, and its atrophy is strongly associated to memory deficits, a hallmark of the disease. In the discrimination tasks, hippocampal volumes are effective at distinguishing between CN, MCI, and AD groups, which aligns with the well-established literature on the role of the hippocampus in cognitive decline [26]. The early inclusion of hippocampal features in both age and discrimination-ordered sets further underscores its dual importance in aging and AD progression. This is consistent with previous findings showing that hippocampal connectivity is indeed significantly affected by aging,

but, other circuits—such as the fronto-striato-thalamic network—are even more severely disrupted (see Fig. S4 in Bonifazi et al) [27]. Thus, although the hippocampus is involved in both aging and AD, its prominent role in AD disease may reflect a stronger association with pathological processes rather than normal aging, highlighting the interaction between these two mechanisms.

Our results show clear distinctions between BrainAge models trained on feature sets ranked by age than those ranked by their discriminative power across clinical groups. When features are ranked based on their relationship with age, the MAE of BrainAge models begin at a lower value, indicating that the model achieves higher age prediction accuracy with the top-ranked features. In this case, when using the produced delta for classification the AUC increases steadily as more features are included in the BrainAge model. Thus, reflecting progressively improved ability of the generated deltas to discriminate between clinical groups. This suggests that age-related features provide a solid foundation for accurate age prediction, while classification performance improves progressively as additional, more diverse features are incorporated.

Conversely, when features are ranked by their discriminative power, the BrainAge model's delta AUC shows a higher intercept, reflecting better initial classification performance. However, the MAE of the BrainAge model has a higher intercept compared to the age-ranked feature sets, indicating less accurate age predictions at the onset. Hence, better age prediction does not always mean better classification power. A key transition point occurs when neuroimaging features start to appear in the discrimination-ordered set, after which the AUC stabilizes and MAE decreases. However, we do not see an increase in AUC. This indicates that neuroimaging features enhance age prediction but worsen classification accuracy.

Our previous study showed that models just trained on neuropsychological data outperformed those trained on neuroimaging data on the classification task but not in age prediction error [12].

We acknowledge that this is to be expected since neuropsychological testing is used to diagnose and is an inherent part of the process. However, as argued above this is not circular reasoning since the BrainAge models are trained in the healthy cohort and never use neuropsychological data from MCI or AD during training. Hence, it is important to determine what the use of the calculated deltas will be when creating BrainAge models. Subsequently, the model should be trained with the features required for its respective task.

Recent studies have emphasized the importance of distinguishing between brain changes due to typical aging and those caused by AD. Hwang et al. (2022) introduced machine learning models such as SPARE-AD and SPARE-BA to decouple the effects of aging and neurodegeneration in brain imaging data [28]. Their results showed that by employing conservative molecular diagnoses and introducing Alzheimer's continuum cases, it was possible to derive more specific neuroanatomical biomarkers for aging and AD, reducing the overlap in brain regions affected by both processes. Similarly, Binette et al. (2020) investigated grey matter differences across a wide age range and in AD patients, finding that while both aging and AD contribute to widespread brain atrophy, AD uniquely disrupts whole-brain morphometric organization [29]. This disruption in grey matter pattern, more than volume loss, was strongly associated with cognitive impairment, highlighting the need to account for these distinct mechanisms when assessing BrainAge. Alternatively, feature importance for biological aging FIBA has been proposed as a method to improve the relevance of age models by identifying features that contribute specifically to biological rather than chronological age, refining their association with cognitive traits [30].

Our results indicate that training a logistic regression directly on the features either outperforms or equals a logistic regression trained on BrainAge deltas. The results tend to be similar when the classification task is harder, like when distinguishing sMCI vs pMCI. This is in line with a recent

study that found that neural network-based BrainAge models, when retrained for classification, perform worse than models trained directly for classification [31]. Training on different regression models, such as linear regression ridge regression, and SVM does not improve performance. This outcome is likely influenced by the high linear correlation among neuroimaging features, as a model using a single key feature often performs similarly to those incorporating multiple features. Gray matter volume in particular plays a dominant role in both age prediction and discrimination across clinical groups, highlighting its central importance in distinguishing disease states.

However, BrainAge offers an advantage over training a logistic regressor directly with features. BrainAge deltas encapsulate disease progression within a single metric rather than requiring training separate classification models for each clinical group. Prior studies have shown that BrainAge deltas correlate with the time to conversion from MCI to AD [12], making them a potentially useful quantitative biomarker for assessing risk. This further underscores the need for careful consideration when applying BrainAge models to clinical decision-making.

Adding age or BrainAge delta as additional features does not consistently improve classification accuracy, as indicated by minimal AUC increases that remain within standard deviations. This variability suggests that neither age nor delta provides substantial independent discriminative value beyond the primary neuroimaging features. Notably, repeating the classification process with different data partitions and cross-validation folds leads to wide fluctuations in results, emphasizing the sensitivity of the models to sample variation. This finding points to the need for further research to better understand the potential utility of age and delta features in classification, as well as the robustness of these models across varying sample conditions.

Conclusion

Our findings demonstrate that neuroimaging and neuropsychological features play distinct roles in age prediction and disease classification in the context of BrainAge modeling, with neuroimaging features excelling in age prediction and neuropsychological features showing greater sensitivity to clinical distinctions. The hippocampus and grey matter volume emerge as critical biomarkers in both aging and Alzheimer's disease progression. The instability of results across different model configurations and data splits suggests a need for further exploration into feature robustness and optimal model selection based on task-specific requirements. Overall, while BrainAge deltas can provide a useful single metric associated with the risk of developing Alzheimer's disease, their application should be carefully considered in the context of specific clinical and research goals.

Ethic statement

Not applicable. Data was obtained from the Alzheimer's Neuroimaging Data Initiative (ADNI) see www.adni.loni.usc.edu.

CRediT authorship

Jorge Garcia Condado: Conceptualization, Methodology, Data curation, Formal analysis, Visualization, Writing – original draft, Writing – review & editing. **Ines Verdugo Recuerdo:** Methodology, Formal analysis, Visualization, Writing – review & editing. **Iñigo Tellaetxe Elorriaga**: Data curation, Visualization, Writing – original draft **Colin Birkhenbil:** Writing – review & editing. **Maria Carrigan**: Writing – review & editing **Ibai Diez**: Writing – review & editing. **Rachel F. Buckley**: Writing – review & editing. **Asier Erramuzpe:** Supervision, Project administration, Writing – review & editing. **Jesus M. Cortes**: Supervision, Project administration, Writing – review & editing.

Declaration of competing interest

None declared

Acknowledgments

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 the analysis or the writing of this report. A complete listing of ADNI investigators can be found here.

We want to thank Vanessa Goméz Verdejo from Universidad Carlos III for her valuable comments on the work.

The project that gave rise to these results received the support of a fellowship from "la Caixa" Foundation (ID 100010434) to J.G.C. The fellowship code is LCF/BQ/DI21/11860030. J.M.C acknowledges financial support from the Health Department of the Basque Country (grants 202211103 and 2023111002). A.E. is supported by the Spanish Ministry of Science and Innovation, grant RYC2021-032390-I. J.M.C. and A.E. are supported by Ikerbasque: The Basque Foundation for Science and the Spanish Ministry of Science under Grant PID2023-148012OB-I00.

Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug

Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

Code Availability

The software used in this work is open source and publicly available: www.github.com/compneurobilbao/ageml.

References

[1] K. Franke, G. Ziegler, S. Klöppel, C. Gaser, Estimating the age of healthy subjects from T1-weighted MRI scans using kernel methods: Exploring the influence of various parameters, NeuroImage 50 (2010) 883–892. https://doi.org/10.1016/j.neuroimage.2010.01.005.

[2] C. Gaser, K. Franke, S. Klöppel, N. Koutsouleris, H. Sauer, Alzheimer's Disease Neuroimaging Initiative, BrainAGE in Mild Cognitive Impaired Patients: Predicting the Conversion to Alzheimer's Disease, PLoS ONE 8 (2013) e67346. https://doi.org/10.1371/journal.pone.0067346.

[3] L. Baecker, R. Garcia-Dias, S. Vieira, C. Scarpazza, A. Mechelli, Machine learning for brain age prediction: Introduction to methods and clinical applications, eBioMedicine 72 (2021) 103600. https://doi.org/10.1016/j.ebiom.2021.103600.

[4] S. Moguilner, S. Baez, H. Hernandez, J. Migeot, A. Legaz, R. Gonzalez-Gomez, F.R. Farina, P. Prado, J. Cuadros, E. Tagliazucchi, F. Altschuler, M.A. Maito, M.E. Godoy, J. Cruzat, P.A. Valdes-Sosa, F. Lopera, J.F. Ochoa-Gómez, A.G. Hernandez, J. Bonilla-Santos, R.A. Gonzalez-Montealegre, R. Anghinah, L.E. d'Almeida Manfrinati, S. Fittipaldi, V. Medel, D. Olivares, G.G. Yener, J. Escudero, C. Babiloni, R. Whelan, B. Güntekin, H. Yırıkoğulları, H. Santamaria-Garcia, A.F. Lucas, D. Huepe, G. Di Caterina, M. Soto-Añari, A. Birba, A. Sainz-Ballesteros, C. Coronel-Oliveros, A. Yigezu, E. Herrera, D. Abasolo, K. Kilborn, N. Rubido, R.A. Clark, R. Herzog, D. Yerlikaya, K. Hu, M.A. Parra, P. Reyes, A.M. García, D.L. Matallana, J.A. Avila-Funes, A. Slachevsky, M.I. Behrens, N. Custodio, J.F. Cardona, P. Barttfeld, I.L. Brusco, M.A. Bruno, A.L. Sosa Ortiz, S.D. Pina-Escudero, L.T. Takada, E. Resende, K.L. Possin, M.O. De Oliveira, A. Lopez-Valdes, B. Lawlor, I.H. Robertson, K.S. Kosik, C. Duran-Aniotz, V. Valcour, J.S. Yokoyama, B. Miller, A. Ibanez, Brain clocks capture diversity and disparities in aging and dementia across geographically diverse populations, Nat Med (2024). https://doi.org/10.1038/s41591-024-03209-x.

[5] A. Puc, V. Struc, K. Grm, Analysis of Race and Gender Bias in Deep Age Estimation Models, in: 2020 28th European Signal Processing Conference (EUSIPCO), IEEE, Amsterdam, Netherlands, 2021: pp. 830–834. https://doi.org/10.23919/Eusipco47968.2020.9287219.

[6] K. Franke, C. Gaser, Ten Years of BrainAGE as a Neuroimaging Biomarker of Brain Aging: What Insights Have We Gained?, Front. Neurol. 10 (2019) 789. https://doi.org/10.3389/fneur.2019.00789.

[7] A.G. de Lange, M. Anatürk, J. Rokicki, L.K.M. Han, K. Franke, D. Alnæs, K.P. Ebmeier, B. Draganski, T. Kaufmann, L.T. Westlye, T. Hahn, J.H. Cole, Mind the gap: Performance metric evaluation in brain-age prediction, Human Brain Mapping 43 (2022) 3113–3129. https://doi.org/10.1002/hbm.25837.

[8] Y. Yu, H. Cui, S.S. Haas, F. New, N. Sanford, K. Yu, D. Zhan, G. Yang, J. Gao, D. Wei, J. Qiu, N. Banaj, D.I. Boomsma, A. Breier, H. Brodaty, R.L. Buckner, J.K. Buitelaar, D.M. Cannon, X. Caseras, V.P. Clark, P.J. Conrod, F. Crivello, E.A. Crone, U. Dannlowski, C.G. Davey, L. de Haan, G.I. de Zubicaray, A. Di Giorgio, L. Fisch, S.E. Fisher, B. Franke, D.C. Glahn, D. Grotegerd, O. Gruber, R.E. Gur, R.C. Gur, T. Hahn, B.J. Harrison, S. Hatton, I.B. Hickie, H.E. Hulshoff Pol, A.J. Jamieson, T.L. Jernigan, J. Jiang, A.J. Kalnin, S. Kang, N.A. Kochan, A. Kraus, J. Lagopoulos, L. Lazaro, B.C. McDonald, C. McDonald, K.L. McMahon, B. Mwangi, F. Piras, R. Rodriguez-Cruces, J. Royer, P.S. Sachdev, T.D. Satterthwaite, A.J. Saykin, G. Schumann, P. Sevaggi, J.W. Smoller, J.C. Soares, G. Spalletta, C.K. Tamnes, J.N. Trollor, D. Van't Ent, D. Vecchio, H. Walter, Y. Wang, B. Weber, W. Wen, L.M. Wierenga, S.C.R. Williams, M. Wu, G.B. Zunta-Soares, B. Bernhardt, P. Thompson, S. Frangou, R. Ge, Brain-age prediction: Systematic evaluation of site effects, and sample age range and size, Hum Brain Mapp 45 (2024) e26768. https://doi.org/10.1002/hbm.26768.

[9] M. Korbmacher, M.-Y. Wang, R. Eikeland, R. Buchert, E. Leonardsen, L.T. Westlye, I.I. Maximov, K. Specht, Brain age predictions in longitudinal data reveal the importance of scan quality and field strength, Neuroscience, 2023. https://doi.org/10.1101/2023.03.31.535038.

[10] R.K. Gebre, M.L. Senjem, S. Raghavan, C.G. Schwarz, J.L. Gunter, E.I. Hofrenning, R.I. Reid, K. Kantarci, J. Graff-Radford, D.S. Knopman, R.C. Petersen, C.R. Jack, P. Vemuri,

Cross–scanner harmonization methods for structural MRI may need further work: A comparison study, NeuroImage 269 (2023) 119912. https://doi.org/10.1016/j.neuroimage.2023.119912.
[11] L. Dular, Ž. Špiclin, BASE: Brain Age Standardized Evaluation, NeuroImage 285 (2024) 120469. https://doi.org/10.1016/j.neuroimage.2023.120469.
[12] J. Garcia Condado, J.M. Cortes, for the Alzheimer's Disease Neuroimaging Initiative, NeuropsychBrainAge: A biomarker for conversion from mild cognitive impairment to Alzheimer's disease, Alz & Dem Diag Ass & Dis Mo 15 (2023) e12493. https://doi.org/10.1002/dad2.12493.
[13] J.G. Condado, I.T. Elorriaga, J.M. Cortes, A. Erramuzpe, AgeML: Age modeling with Machine Learning, IEEE J. Biomed. Health Inform. (2025) 1–11. https://doi.org/10.1109/JBHI.2025.3531017.
[14] R.C. Petersen, P.S. Aisen, L.A. Beckett, M.C. Donohue, A.C. Gamst, D.J. Harvey, C.R. Jack, W.J. Jagust, L.M. Shaw, A.W. Toga, J.Q. Trojanowski, M.W. Weiner, Alzheimer's Disease Neuroimaging Initiative (ADNI): Clinical characterization, Neurology 74 (2010) 201–209. https://doi.org/10.1212/WNL.0b013e3181cb3e25.
[15] S.M. Smith, Y. Zhang, M. Jenkinson, J. Chen, P.M. Matthews, A. Federico, N. De Stefano, Accurate, Robust, and Automated Longitudinal and Cross-Sectional Brain Change Analysis, NeuroImage 17 (2002) 479–489. https://doi.org/10.1006/nimg.2002.1040.
[16] S.M. Smith, M. Jenkinson, M.W. Woolrich, C.F. Beckmann, T.E.J. Behrens, H. Johansen-Berg, P.R. Bannister, M. De Luca, I. Drobnjak, D.E. Flitney, R.K. Niazy, J. Saunders, J. Vickers, Y. Zhang, N. De Stefano, J.M. Brady, P.M. Matthews, Advances in functional and structural MR image analysis and implementation as FSL, NeuroImage 23 (2004) S208–S219. https://doi.org/10.1016/j.neuroimage.2004.07.051.
[17] B. Patenaude, S.M. Smith, D.N. Kennedy, M. Jenkinson, A Bayesian model of shape and appearance for subcortical brain segmentation, NeuroImage 56 (2011) 907–922. https://doi.org/10.1016/j.neuroimage.2011.02.046.
[18] M.F. Folstein, S.E. Folstein, P.R. McHugh, "Mini-mental state," Journal of Psychiatric Research 12 (1975) 189–198. https://doi.org/10.1016/0022-3956(75)90026-6.
[19] A new rating scale for Alzheimer's disease, AJP 141 (1984) 1356–1364. https://doi.org/10.1176/ajp.141.11.1356.
[20] R.I. Pfeffer, T.T. Kurosaki, C.H. Harrah, J.M. Chance, S. Filos, Measurement of Functional Activities in Older Adults in the Community, Journal of Gerontology 37 (1982) 323–329. https://doi.org/10.1093/geronj/37.3.323.
[21] Z.S. Nasreddine, N.A. Phillips, V. BÃ©dirian, S. Charbonneau, V. Whitehead, I. Collin, J.L. Cummings, H. Chertkow, The Montreal Cognitive Assessment, MoCA: A Brief Screening Tool For Mild Cognitive Impairment: MOCA: A BRIEF SCREENING TOOL FOR MCI, Journal of the American Geriatrics Society 53 (2005) 695–699. https://doi.org/10.1111/j.1532-5415.2005.53221.x.
[22] for the Alzheimer's Disease Neuroimaging Initiative, P.K. Crane, A. Carle, L.E. Gibbons, P. Insel, R.S. Mackin, A. Gross, R.N. Jones, S. Mukherjee, S.M. Curtis, D. Harvey, M. Weiner, D. Mungas, Development and assessment of a composite score for memory in the Alzheimer's Disease Neuroimaging Initiative (ADNI), Brain Imaging and Behavior 6 (2012) 502–516. https://doi.org/10.1007/s11682-012-9186-z.
[23] for the Alzheimer's Disease Neuroimaging Initiative, L.E. Gibbons, A.C. Carle, R.S. Mackin, D. Harvey, S. Mukherjee, P. Insel, S.M. Curtis, D. Mungas, P.K. Crane, A composite

score for executive functioning, validated in Alzheimer's Disease Neuroimaging Initiative (ADNI) participants with baseline mild cognitive impairment, Brain Imaging and Behavior 6 (2012) 517–527. https://doi.org/10.1007/s11682-012-9176-1.

[24] Scikit-learn: Machine Learning in Python, 2011 12 (n.d.) 2825–2830.

[25] A.-M.G. de Lange, J.H. Cole, Commentary: Correction procedures in brain-age prediction, NeuroImage: Clinical 26 (2020) 102229. https://doi.org/10.1016/j.nicl.2020.102229.

[26] Y. Mu, F.H. Gage, Adult hippocampal neurogenesis and its role in Alzheimer's disease, Mol Neurodegeneration 6 (2011) 85. https://doi.org/10.1186/1750-1326-6-85.

[27] P. Bonifazi, A. Erramuzpe, I. Diez, I. Gabilondo, M.P. Boisgontier, L. Pauwels, S. Stramaglia, S.P. Swinnen, J.M. Cortes, Structure–function multi-scale connectomics reveals a major role of the fronto-striato-thalamic circuit in brain aging, Human Brain Mapping 39 (2018) 4663–4677. https://doi.org/10.1002/hbm.24312.

[28] G. Hwang, A. Abdulkadir, G. Erus, M. Habes, R. Pomponio, H. Shou, J. Doshi, E. Mamourian, T. Rashid, M. Bilgel, Y. Fan, A. Sotiras, D. Srinivasan, J.C. Morris, M.S. Albert, N.R. Bryan, S.M. Resnick, I.M. Nasrallah, C. Davatzikos, D.A. Wolk, from the iSTAGING consortium, for the ADNI, Disentangling Alzheimer's disease neurodegeneration from typical brain ageing using machine learning, Brain Communications 4 (2022) fcac117. https://doi.org/10.1093/braincomms/fcac117.

[29] A. Pichet Binette, J. Gonneaud, J.W. Vogel, R. La Joie, P. Rosa-Neto, D.L. Collins, J. Poirier, J.C.S. Breitner, S. Villeneuve, E. Vachon-Presseau, for the Alzheimer's Disease Neuroimaging Initiative, the PREVENT-AD Research Group, Morphometric network differences in ageing versus Alzheimer's disease dementia, Brain 143 (2020) 635–649. https://doi.org/10.1093/brain/awz414.

[30] H.S.-H. Oh, J. Rutledge, D. Nachun, R. Pálovics, O. Abiose, P. Moran-Losada, D. Channappa, D.Y. Urey, K. Kim, Y.J. Sung, L. Wang, J. Timsina, D. Western, M. Liu, P. Kohlfeld, J. Budde, E.N. Wilson, Y. Guen, T.M. Maurer, M. Haney, A.C. Yang, Z. He, M.D. Greicius, K.I. Andreasson, S. Sathyan, E.F. Weiss, S. Milman, N. Barzilai, C. Cruchaga, A.D. Wagner, E. Mormino, B. Lehallier, V.W. Henderson, F.M. Longo, S.B. Montgomery, T. Wyss-Coray, Organ aging signatures in the plasma proteome track health and disease, Nature 624 (2023) 164–172. https://doi.org/10.1038/s41586-023-06802-1.

[31] T.W.K. Tan, K.-N. Nguyen, C. Zhang, R. Kong, S.F. Cheng, F. Ji, J.S.X. Chong, E.J.Y. Chong, N. Venketasubramanian, C. Orban, M.W.L. Chee, C. Chen, J.H. Zhou, B.T.T. Yeo, Alzheimer's Disease Neuroimaging Initiative, Australian Imaging Biomarkers and Lifestyle Study of Aging, Evaluation of Brain Age as a Specific Marker of Brain Health, (2024). https://doi.org/10.1101/2024.11.16.623903.

Figures

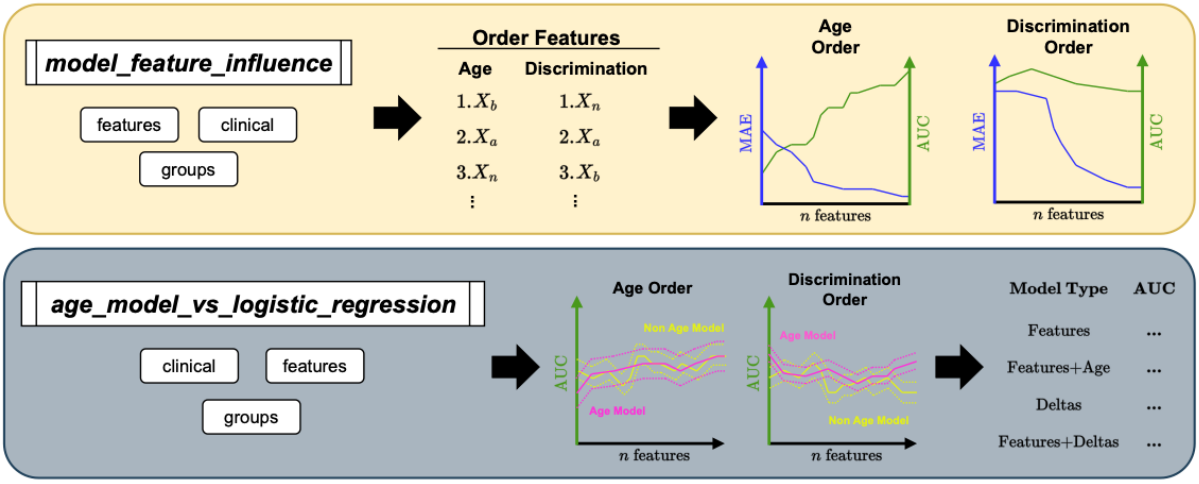


Figure 1. Overview of the two new commands for AgeML [13] to explore the relationship between age prediction and classification accuracy. *Top panel:* Model Feature Influence pipeline schematic. First, the provided features are ranked according to their mutual information with age or according to their discriminative power to classify the specified clinical groups. Then, age regression and clinical classification models are trained with the computed orderings to evaluate how the progressive addition of features affects the performance of the models. The progression curves are automatically plotted. *Bottom panel:* Age Models versus direct logistic regression pipeline. The given features are first ranked according to the same criterion from above. After, two logistic regressor types are trained to classify the specified clinical groups; one based on an age delta computed from Age models; and the other directly using the features. The performance of the classifiers is plotted trained with increasingly more features added in the computed orderings, similar to the panel above. A summary table and the progression curves are automatically output.

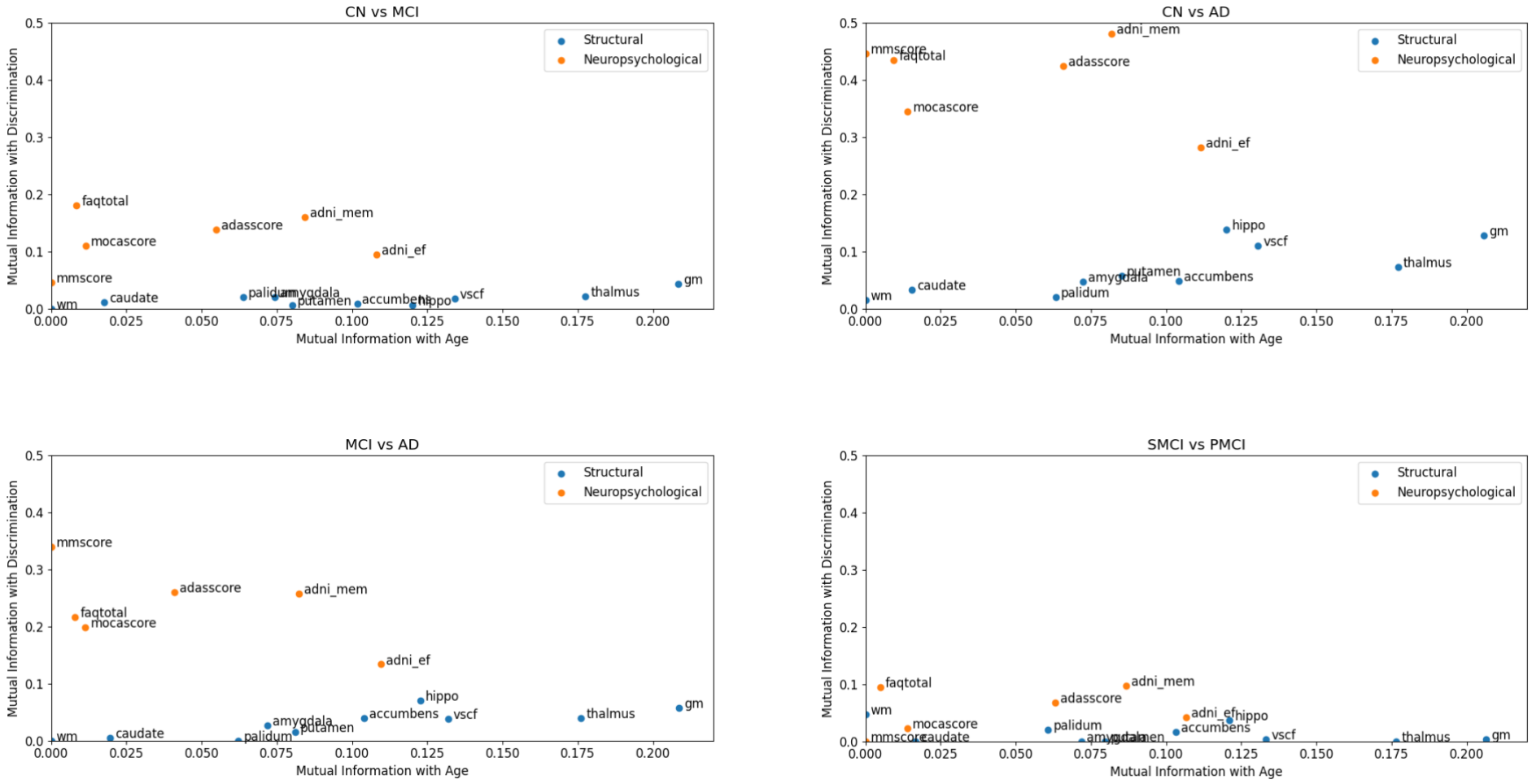


Figure 2. Mutual information of each feature with Age and their discriminative power. In blue are features derived from neuroimaging metrics and in orange features derived from neuropsychological testis. Control (CN), Alzheimer's Disease (AD), Mild Cognitive Impairment (MCI), stable MCI (sMCI), and progressive MCI (pMCI)

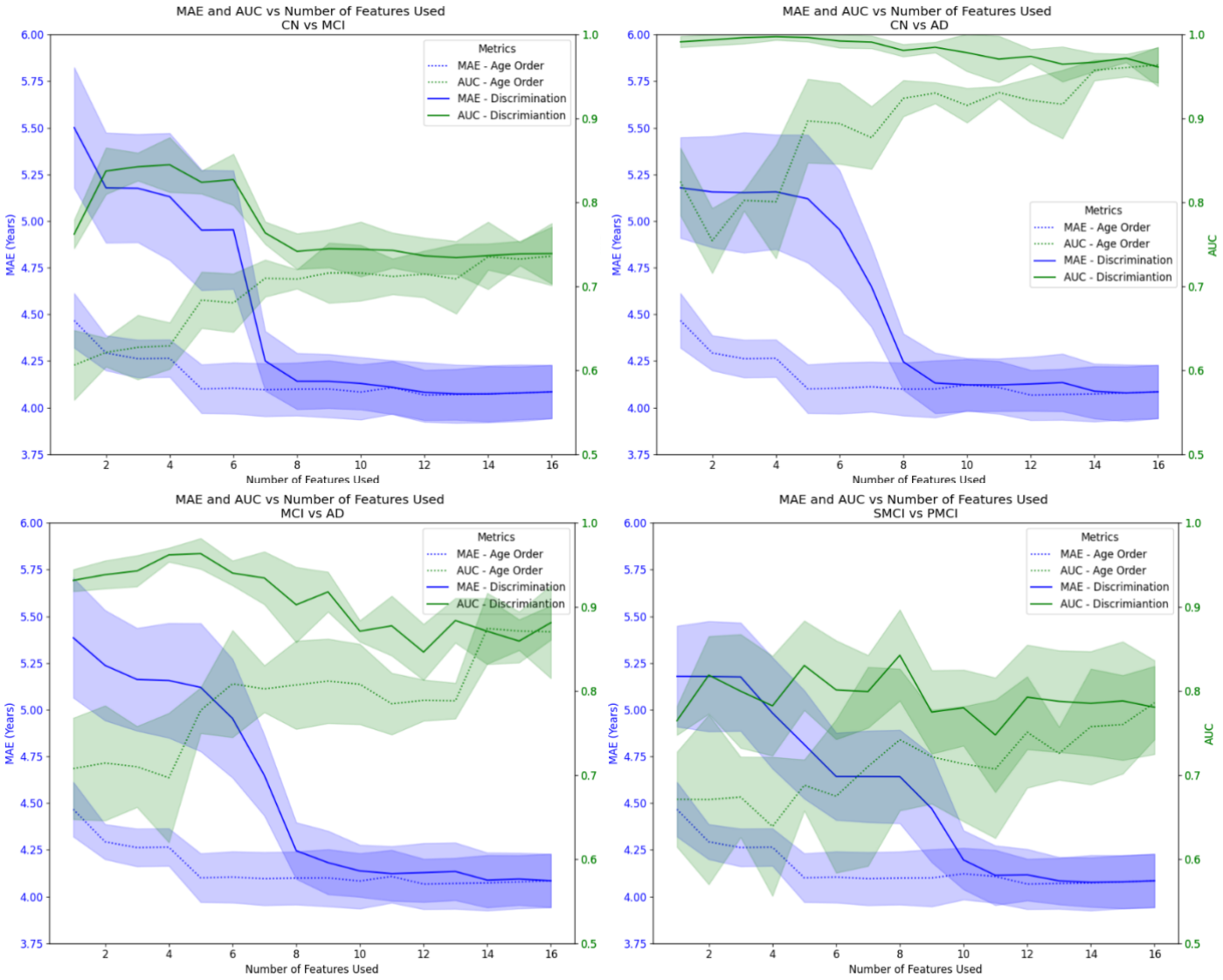
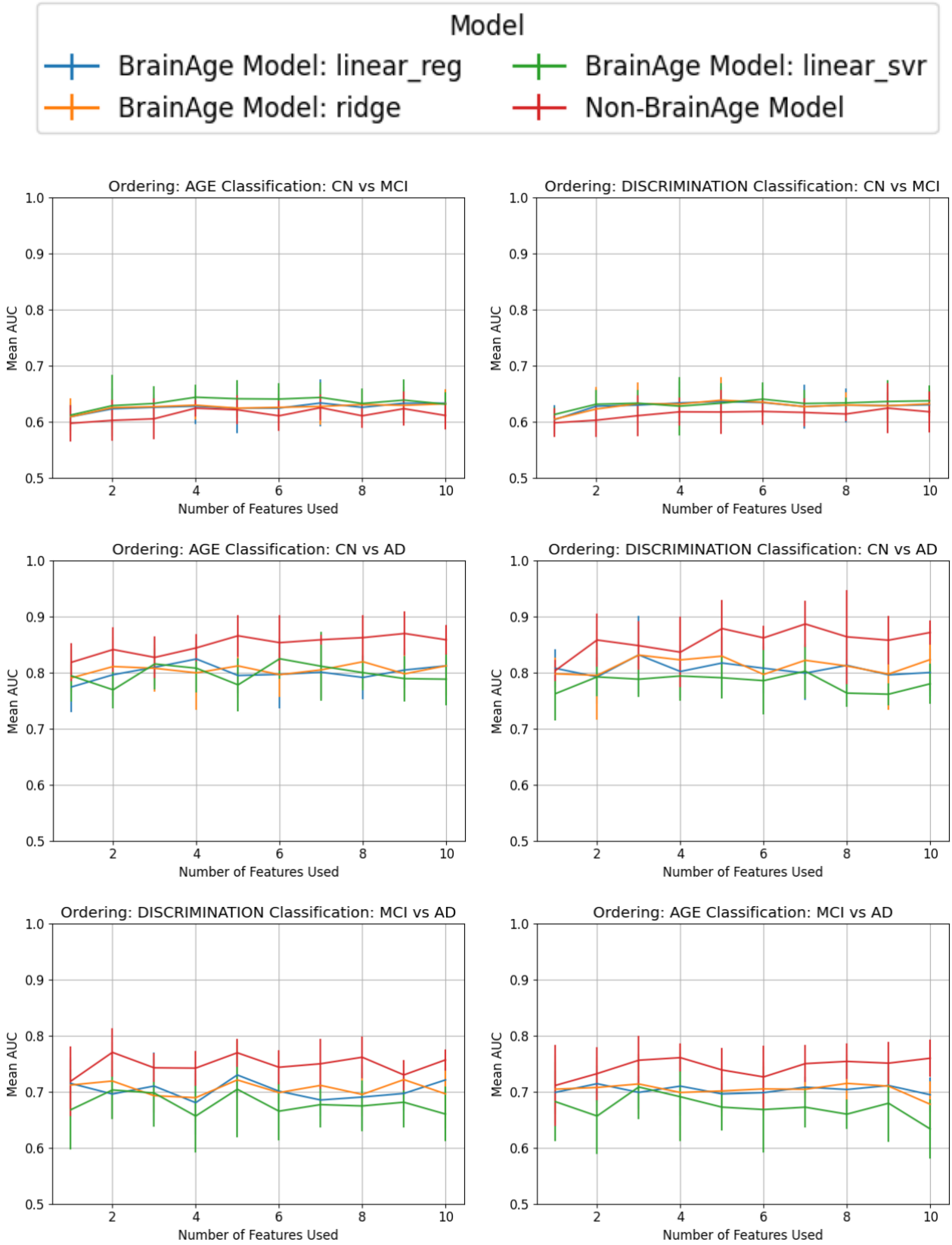


Figure. 3. Comparison of the performance metrics, Mean Absolute Error (MAE), and Area Under the Curve (AUC) for different health condition groups using BrainAge Modelling. Features are added to the BrainAge model in descending order based on their age relationship (dotted line). Additionally, features are added to the BrainAge model in descending order according to their importance in discriminating between the following groups: Control (CN), Alzheimer's Disease (AD), Mild Cognitive Impairment (MCI), stable MCI (sMCI), and progressive MCI (pMCI) (solid line).


Model
BrainAge Model: linear_reg
BrainAge Model: linear_svr
BrainAge Model: ridge
Non-BrainAge Model
Ordering: AGE Classification: CN vs MCI
Ordering: DISCRIMINATION Classification: CN vs MCI
Ordering: AGE Classification: CN vs AD
Ordering: DISCRIMINATION Classification: CN vs AD
Ordering: DISCRIMINATION Classification: MCI vs AD
Ordering: AGE Classification: MCI vs AD
Mean AUC
Number of Features Used
1.0
0.9
0.8
0.7
0.6
0.5
2
4
6
8
10

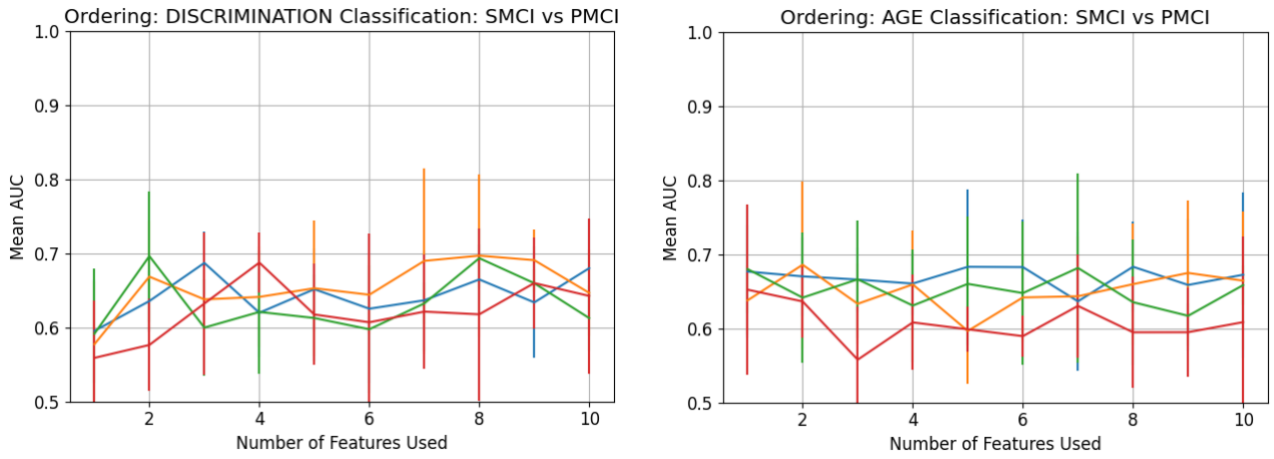


Figure. 4. Comparison of the Area Under the Curve (AUC) using different machine learning models for classification using neuroimaging features. *Blue line:* Input to logistic regressor: Delta. BrainAge model: linear regression. *Orange line:* Input to logistic regressor: Delta. BrainAge model: Ridge. *Green line:* Input to logistic regressor: Delta. BrainAge model: Support Vector Regressor. *Red line:* Input to logistic regressor: Neuroimaging Features, No BrainAge modeling. Models are tested and trained across clinical classification groups and features ordering by Age Relationship and by discrimination between groups order (Control Group (CN), Alzheimer Disease (AD), Mild Cognitive Impairment (MCI), stable Mild Cognitive Impairment (sMCI) and progressive Mild Cognitive Impairment (pMCI)).

Tables

Table 1. Cohort demographics

Characteristic	CN N = 629[1]	MCI N = 635[1]	AD N = 208[1]	sMCI N = 238[1]	pMCI N = 98[1]	p-value[2]
Gender						<0.001
Female	364 (58%)	279 (44%)	87 (42%)	110 (46%)	43 (44%)	
Male	265 (42%)	356 (56%)	121 (58%)	128 (54%)	55 (56%)	
Age (Years)	72 (7)	72 (8)	75 (8)	73 (8)	73 (7)	<0.001
Education (Years)	17 (2)	16 (3)	16 (3)	16 (3)	16 (3)	<0.001

[1]n (%); Mean (SD)

[2]Pearson's Chi-squared test; Kruskal-Wallis rank sum test

Table 2. Results of ranking features across different orderings. The features are ordered in descending importance according to the variable indicated in the column header. The firs column indicates ordering according to mutual information with age, while the subsequent columns make a comparison between the mutual information of features for the following groups: Control Group (CN), Alzheimer's Diseases (AD), Mild Cognitive Impairment (MCI), Stable Mild Cognitive Impairment (sMCI) and Progressive Mild Cognitive Impairment (pMCI)

Set	Age	CN vs AD	CN vs MCI	MCI vs AD	sMCI vs pMCI
1	Gray Matter	ADNI Memory	FAQ	MMSE	FAQ
2	Thalamus	MMSE	ADNI Memory	ADAS	ADAS

3	VSCF	FAQ	ADAS	ADNI Memory	ADNI Memory
4	Hippocampus	ADAS	MoCA	FAQ	White Matter
5	ADNI Executive Function	MoCA	ADNI Executive Function	MoCA	ADNI Executive Function
6	Accumbens	ADNI Executive Function	MMSE	ADNI Executive Function	Hippocampus
7	ADNI Memory	Hippocampus	Gray Matter	Hippocampus	Palidum
8	Putamen	Gray Matter	Thalamus	Gray Matter	Accumbens
9	Amygadala	CSVF	Amygdala	Accumbens	MMSE
10	Palidum	Thalamus	Palidum	Thalamus	MoCA
11	ADAS	Putamen	VSCF	VSCF	Gray Matter
12	Caudate	Accumbens	Caudate	Amygdala	VSCF
13	MoCA	Amygda;a	Accumbens	Putamen	Thalamus
14	FAQ	Caudate	Putamen	Caudate	Caudate
15	White Matter	Palidum	Hippocampus	White Matter	Putamen
16	MMSE	White Matter	White Matter	Palidum	Amygdala

Table 3. Comparison of the Area Under the Curve (AUC) across various clinical classifications using different input feature sets. The classification was performed using a Logistic Regressor and included all brain structural features for the following groups: Control Group (CN), Alzheimer's Disease (AD), Mild Cognitive Impairment (MCI), stable Mild Cognitive Impairment (sMCI), and progressive Mild Cognitive Impairment (pMCI).

Groups	Features	Features + Age	Delta	Features + Delta
CN vs MCI	0.62 ± 0.02	0.64 ± 0.03	0.63 ± 0.05	0.63 ± 0.03
CN vs AD	0.85 ± 0.03	0.85 ± 0.03	0.80 ± 0.06	0.88 ± 0.06
MCI vs AD	0.78 ± 0.03	0.76 ± 0.06	0.71 ± 0.05	0.73 ± 0.04
pMCI vs sMCI	0.60 ± 0.10	0.68 ± 0.06	0.64 ± 0.09	0.65 ± 0.10