Medical Image Analysis 90 (2023) 102913

Contents lists available at ScienceDirect

Medical Image Analysis

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

Gaussian Process-based prediction of memory performance and biomarker status in ageing and Alzheimer's disease—A systematic model evaluation

A. Nemali [1,2,*], N. Vockert [2], D. Berron [2], A. Maas [2], J. Bernal [2], R. Yakupov [2], O. Peters [4,5], D. Gref [5], N. Cosma [5], L. Preis [5], J. Priller [4,6,28,29], E. Spruth [4,6], S. Altenstein [4,6], A. Lohse [6], K. Fliessbach [7,9], O. Kimmich [7], I. Vogt [7], J. Wiltfang [11,12,26], N. Hansen [12], C. Bartels [12], B.H. Schott [11,12,3], F. Maier [13], D. Meiberth [13], W. Glanz [1], E. Incesoy [1,2], M. Butryn [2], K. Buerger [14,15], D. Janowitz [15], R. Pernecky [14,20,21,22], B. Rauchmann [20], L. Burow [20], S. Teipel [16,17], I. Kilimann [16,17], D. Göerß [17], M. Dyrba [16], C. Laske [18,19], M. Munk [18,31], C. Sanzenbacher [18], S. Müller [31], A. Spottke [7,10], N. Roy [7], M. Heneka [7,8], F. Brosseron [7], S. Roeske [7], L. Dobisch [2], A. Ramirez [7,10,25,27,30], M. Ewers [14,15], P. Dechent [23], K. Scheffler [24], L. Kleineidam [9], S. Wolfsgruber [7,9], M. Wagner [7,9], F. Jessen [7,13,25], E. Duzel [1,2,a], G. Ziegler [1,2,a]

[1] *Institute of Cognitive Neurology and Dementia Research, Otto-von-Guericke University Magdeburg, Germany*
[2] *German Center for Neurodegenerative Diseases (DZNE), Magdeburg, Germany*
[3] *Leibniz Institute for Neurobiology, Magdeburg, Germany*
[4] *German Center for Neurodegenerative Diseases (DZNE), Berlin, Germany*
[5] *Charité-Universitätsmedizin Berlin, Campus Benjamin Franklin, Department of Psychiatry, Hindenburgdamm 30, 12203, Berlin, Germany*
[6] *Department of Psychiatry and Psychotherapy, Charité, Charitéplatz 1, 10117 Berlin, Germany*
[7] *German Center for Neurodegenerative Diseases (DZNE), Bonn, Venusberg-Campus 1, 53127 Bonn, Germany*
[8] *Department of Psychiatry and Psychotherapy, University of Bonn, Venusberg-Campus 1, 53127 Bonn, Germany*
[9] *University of Bonn Medical Center, Department of Neurodegenerative Disease and Geriatric Psychiatry/Psychiatry, Venusberg-Campus 1, 53127 Bonn, Germany*
[10] *Department of Neurology, University of Bonn, Venusberg-Campus 1, 53127 Bonn, Germany*
[11] *German Center for Neurodegenerative Diseases (DZNE), Goettingen, Germany*
[12] *Department of Psychiatry and Psychotherapy, University Medical Center Goettingen, University of Goettingen, Von-Siebold-Str. 5, 37075 Goettingen, Germany*
[13] *Department of Psychiatry, University of Cologne, Medical Faculty, Kerpener Strasse 62, 50924 Cologne, Germany*
[14] *German Center for Neurodegenerative Diseases (DZNE, Munich), Feodor-Lynen-Strasse 17, 81377 Munich, Germany*
[15] *Institute for Stroke and Dementia Research (ISD), University Hospital, LMU Munich, Feodor-Lynen-Strasse 17, 81377 Munich, Germany*
[16] *German Center for Neurodegenerative Diseases (DZNE), Rostock, Germany*
[17] *Department of Psychosomatic Medicine, Rostock University Medical Center, Gehlsheimer Str. 20, 18147 Rostock, Germany*
[18] *German Center for Neurodegenerative Diseases (DZNE), Tübingen, Germany*
[19] *Section for Dementia Research, Hertie Institute for Clinical Brain Research and Department of Psychiatry and Psychotherapy, University of Tübingen, Tübingen, Germany*
[20] *Department of Psychiatry and Psychotherapy, University Hospital, LMU Munich, Munich, Germany*
[21] *Munich Cluster for Systems Neurology (SyNergy) Munich, Munich, Germany*
[22] *Ageing Epidemiology Research Unit (AGE), School of Public Health, Imperial College London, London, UK*
[23] *MR-Research in Neurosciences, Department of Cognitive Neurology, Georg-August-University Goettingen, Germany*
[24] *Department for Biomedical Magnetic Resonance, University of Tübingen, 72076 Tübingen, Germany*
[25] *Excellence Cluster on Cellular Stress Responses in Aging-Associated Diseases (CECAD), University of Cologne, Joseph-Stelzmann-Strasse 26, 50931 Köln, Germany*
[26] *Neurosciences and Signaling Group, Institute of Biomedicine (iBiMED), Department of Medical Sciences, University of Aveiro, Aveiro, Portugal*
[27] *Division of Neurogenetics and Molecular Psychiatry, Department of Psychiatry and Psychotherapy, Faculty of Medicine and University Hospital Cologne, University of Cologne, Cologne, Germany*
[28] *School of Medicine, Technical University of Munich; Department of Psychiatry and Psychotherapy, Munich, Germany*
[29] *University of Edinburgh and UK DRI, Edinburgh, UK*
[30] *Department of Psychiatry & Glenn Biggs Institute for Alzheimer's and Neurodegenerative Diseases, San Antonio, TX, USA*
[31] *Department of Psychiatry and Psychotherapy, University of Tuebingen, Tuebingen, Germany*

* Corresponding author at: Institute of Cognitive Neurology and Dementia Research, Otto-von-Guericke University Magdeburg, Germany.
E-mail address: aditya.nemali@dzne.de (A. Nemali).

[a] The two authors contributed equally to this paper.

https://doi.org/10.1016/j.media.2023.102913
Received 15 March 2022; Received in revised form 28 June 2023; Accepted 25 July 2023
Available online 14 August 2023

ARTICLE INFO

Keywords:
Brain morphology
Brain ageing
Alzheimer's disease
Gaussian processes
Predictive modeling
Cognitive decline
Classification
Bayesian inference

ABSTRACT

Neuroimaging markers based on Magnetic Resonance Imaging (MRI) combined with various other measures (such as genetic covariates, biomarkers, vascular risk factors, neuropsychological tests etc.) might provide useful predictions of clinical outcomes during the progression towards Alzheimer's disease (AD). The use of multiple features in predictive frameworks for clinical outcomes has become increasingly prevalent in AD research. However, many studies do not focus on systematically and accurately evaluating combinations of multiple input features. Hence, the aim of the present work is to explore and assess optimal combinations of various features for MR-based prediction of (1) cognitive status and (2) biomarker positivity with a multi-kernel learning Gaussian process framework. The explored features and parameters included (A) combinations of brain tissues, modulation, smoothing, and image resolution; (B) incorporating demographics & clinical covariates; (C) the impact of the size of the training data set; (D) the influence of dimensionality reduction and the choice of kernel types. The approach was tested in a large German cohort including 959 subjects from the multicentric longitudinal study of cognitive impairment and dementia (DELCODE). Our evaluation suggests the best prediction of memory performance was obtained for a combination of neuroimaging markers, demographics, genetic information (ApoE4) and CSF biomarkers explaining 57% of outcome variance in out-of-sample predictions. The highest performance for Aβ42/40 status classification was achieved for a combination of demographics, ApoE4, and a memory score while usage of structural MRI further improved the classification of individual patient's pTau status.

1. Introduction

Alzheimer's disease (AD), the most common cause of dementia, is a progressive neurodegenerative disease that affects ageing population worldwide. AD predominantly impairs the memory domain, eventually leading to dementia, in which most cognitive functions are impaired and daily life activities are disrupted (Gaugler et al., 2019). The current understanding of disease pathology suggests that AD progression might start decades before clinical symptoms manifest (Morris, 2005; Jack et al., 2016; Dubois et al., 2016; Beason-Held et al., 2013). According to the well-established amyloid-cascade hypothesis, AD begins with amyloid-beta ($A\beta$) protein accumulation and tau pathology followed by neurodegeneration (atrophy of the neuropil and loss of neurons causing brain atrophy) and cognitive impairment (Murphy and LeVine III, 2010; Jack et al., 2013; Blennow et al., 2012). Neurodegeneration may be characterized by measuring total tau in the CSF using magnetic resonance imaging (MRI) techniques to assess local brain volumes (Frisoni et al., 2010; Knešaurek, 2015; Besson et al., 2015) and by measuring metabolism using positron emission tomography (PET) (Forsberg et al., 2008; Humpel, 2011).

An important challenge in AD research is to identify critical markers for predicting disease severity from various aspects, such as cognitive impairment, structural brain alterations, and biomarkers of amyloid or tau pathology, both cross-sectionally and with regard to longitudinal cognitive decline (Porsteinsson et al., 2021; Li et al., 2021).

Here, we address some of these challenges from the perspective of predictive modeling and machine learning. Machine learning (ML) or artificial intelligence based on computational models and large datasets has boosted medical image analysis and clinical decision support over the past decade (Davatzikos et al., 2019; Mateos-Pérez et al., 2018; Arbabshirani et al., 2017; Salvatore et al., 2016). It is a powerful and promising set of tools to assist doctors with quantitative evidence about individual patients by supporting the diagnosis and prognosis of developmental, psychiatric and neurodegenerative disorders. These brain-based predictions can be beneficial in subjects at risk of dementia to obtain clinical staging and introduce more precise diagnostic and predictive biomarkers of clinical outcomes, which will ultimately lead to an informed choice of potential personalized treatments (Rathore et al., 2017). Two particularly important applications of ML in this context are: First, to implement MR-based predictive models for the presence of brain pathologies (such as amyloid and tau) which are typically measurable via invasive biomarkers (CSF or PET). This might eventually enable the development of less invasive clinical tools for the assessment of potential causes and contributors to cognitive decline. Secondly, the longterm goal is to develop pattern-based prediction models for the downstream consequences of brain pathologies, such as cognitive decline over follow-up visits (i.e. clinical outcome).

Another scientific goal of this approach is to develop decision-support tools that can evaluate brain patterns related to AD severity and patterns that may contribute to cognitive or brain reserves (Stern et al., 2020). Given the regional distribution of amyloid and tau pathologies in AD, the volume of brain regions affected by these pathologies can serve as a predictor of the biomarker burden (Ossenkoppele et al., 2016). To the extent that the same brain regions are also components of neurocognitive networks that enable memory formation, they might predict both biomarker status and cognitive performance cross-sectionally. Moreover, brain (and cognitive) reserve is discussed as aspects of brain structure that vary individually and that moderate the relationship between present pathology (e.g., CSF-biomarkers) and cognitive outcomes (Stern et al., 2020). Distributed morphological patterns used for prediction might reflect both (A) individual protective reserve predispositions as well as (B) the damage induced by protein pathology simultaneously. In case brain (and cognitive) reserve operates the structural MR-based and demographic variables might therefore jointly predict more variance in cognitive performance differences than CSF biomarkers alone.

Studies using regression for the prediction of cognitive performance differences have been introduced to assess cognitive competence/ability and predict IQ in healthy individuals (Rohde and Thompson, 2007; Bradley and Caldwell, 1980; Barber, 2005). Similarly, studies have predicted cognitive performance (Kandiah et al., 2014; Doraiswamy et al., 1998; Woodard et al., 2010) and biomarker status (Prestia et al., 2015; Besson et al., 2015) in older individuals with cognitive impairment. Two aspects of effective predictions are the incorporation of multiple sources of information, such as different morphometric brain properties, for example, from different features or tissue classes (Monté-Rubio et al., 2018) and image modalities (for example,T1 and FLAIR (Amyot et al., 2015; Johnson et al., 2012; Sui et al., 2012), and accounting for demographic background and subject-specific covariates (Ziegler et al., 2014). T1 and FLAIR (Amyot et al., 2015; Johnson et al., 2012; Sui et al., 2012), and accounting for demographic background and subjects-specific covariates (Ziegler et al., 2014).

In recent years, many new tools for predicting the cognitive performance and biomarker status of AD have been integrated to support individualized diagnosis and prognosis (Marquand et al., 2014; Davatzikos et al., 2001; Franke et al., 2010; Rathore et al., 2017). In addition to classical kernel-based methods, such as support vector machines (Shawe-Taylor and Cristianini, 2004), Gaussian processes (Rasmussen and Williams, 2006), multi-kernel approaches (Pettersson-Yeo et al., 2014; Aksman et al., 2019), and neural networks (Fisher et al., 2019; Jo et al., 2019; Dyrba et al., 2021) are increasingly applied for classification and regression.

Machine learning using the so-called kernel method has been shown to have powerful applications in the context of neuroimaging (Dosenbach et al., 2010). A kernel is a bottleneck in learning algorithms that capture (and compress) relevant individual differences, even for high-dimensional input data such as images (Shawe-Taylor et al., 2004). Furthermore, kernel methods have shown promising performance in studies comparing different machine learning frameworks (Jollans et al., 2019), and even similar performance when compared to recent deep neural networks (He et al., 2020). Here, we generalize ideas from Ashburner and Klöppel (2011) for multivariate classification and regression based on (primarily) linear kernels efficiently representing different morphometric brain features (such as gray matter (GM), white matter (WM), and cerebrospinal fluid (CSF); see also Monté-Rubio et al., 2018), and subject-specific covariates (e.g., age, sex, and education; see also Ziegler et al., 2014). Using kernels in combination enables the exploration of the contribution of information from different sources (for instance, morphometric brain features a subject-specific covariates) (Zu et al., 2016). Previous studies have described how to combine multiple kernels (Rakotomamonjy et al., 2008; Gönen and Alpaydın, 2011; Bach et al., 2004) and have reported that the use of multiple kernels can improve the overall predictive performance (Gönen and Alpaydın, 2011; Wilson and Adams, 2013).

There is a compelling argument for a Bayesian treatment of the uncertainties of predictions in clinical and translational applications. In a clinical context, model predictions and their uncertainty estimates are crucial for the decision-making process because they provide clinicians with an informed rationale for using or disregarding the model's prediction. One such class of models that is widely used in the context of neuroimaging is the Gaussian Process (GP) model, which enables a fully probabilistic (Bayesian) prediction of a clinical outcome that provides (posterior) predictive distributions, unlike many alternative approaches. The predictive performance of these models can then be measured using (nested) cross-validation. Additionally, GPs may be used for Bayesian model comparisons by evaluating the marginal log-likelihood (i.e., probability of the outcomes, given the data and hyperparameters) that accounts for both model fit and complexity (Rasmussen, 2006). GPs are flexible tools for the regression of continuous metrics (such as current and memory performance at follow-up) and classification of binary variables (such as biomarker positivity). Moreover, GPs allow the effective handling of multiple information sources of individual differences using multi-kernel learning (Marquand et al., 2014; Rakotomamonjy et al., 2008; Bach et al., 2004; Zu et al., 2016).

In this study, we present an application of multikernel learning to predict the cognitive performance and biomarker status of subjects in a large, well-characterized longitudinal MRI cohort of ageing and AD subjects. The applied GP model for predicting cognitive performance and biomarker status determined the optimal (positive) weighted combination of image-based kernel matrices. Using the optimally selected kernels, we first assess the influence of (A) combinations of brain tissues, modulation, smoothing, and image resolution; (B) incorporating demographics & clinical covariates; (C) the impact of the size of the training data set; (D) the influence of dimensionality reduction and the choice of kernel types; all these aspects are evaluated concerning predictive performance (using the Bayesian model evidence). Finally, the clinical utility of the imaging metrics will be assessed based on their predictive power for new data samples using 10-fold nested cross-validation and longitudinal cognitive follow-up data.

2. Methods

2.1. Predictive model

A predictive model of cognitive ageing accurately predicts a target (or output) variable y^* such as individual memory performance of an older study participant based on a new input test data sample $\mathbf{x}^*$ such as an MRI scan and/or subject-specific covariates. This model can be implemented by learning some unknown function f that maps input data features to outputs using a large set of training data $D = \{\mathbf{X}, \mathbf{y}\}$, where $\mathbf{X}$ represents input data of the training sample of MR features and a set of clinical covariates e.g. $[age, sex, education, ApoE4]$ obtained from concatenating individual features $\mathbf{x}_i$ for all n training subjects. In this work, we focus on two clinically relevant prediction scenarios. First, we aim at an MR-based prediction of individual memory performance (at baseline or follow-up) where all targets y_i (for subject i) refer to observations of a continuous variable. Secondly, using similar MRI feature sets we aim to classify subjects regarding their biomarker status i.e. $y_i \in \{-1, +1\}$. In Appendix A, we briefly revisit technical details about regression of continuous variables and binary classification using Gaussian processes.

2.2. Multi-kernel learning

In order to enable the contribution of multiple imaging modalities (or features) and covariates to the predictions of individual memory performance and bio-marker status, we apply so called multi-kernel learning (MKL) using a linear combination of kernels (Pettersson-Yeo et al., 2014; Aksman et al., 2019; Marquand et al., 2014)

$$\tilde{k}(\mathbf{x}, \mathbf{x}') = \sum_{p=1}^{P} a_p k_p(\mathbf{x}, \mathbf{x}') \tag{1}$$

with a-priori unknown amplitude hyperparameters a_p for up to P features such as GM or WM density images extracted from T1-weighted MRI or simple covariates such as age. Kernels k_p can be chosen using different types (e.g. linear, square exponential (RBF) or automatic relevance determination (ARD) that encode feature similarity differently. The linear combination ofkernels here embeds all subject-specific input data (e.g. voxel-based images) in a high-dimensional space. A key advantage of using the above kernel mappings for GP regression (GPR) and classification (GPC) is to efficiently represent similarities between high-dimensional images such as brain gray matter density patterns in patients (Ashburner and Klöppel, 2011).

2.3. Model optimization and generalization performance

The optimization of the GP models was done by finding the hyperparameters that maximize the marginal log-likelihood using Newton-conjugate gradient descent (Bishop, 2006). To evaluate the out-of-sample predictive performance of the GP models, we used a 10-fold nested cross-validation on the whole dataset, where the outer loop assessed the performance of the model. This strategy was followed to study the influence of different feature combinations, image parameters (smoothness & resolution), and kernels under various feature and model combinations. The hyperparameters of the model were evaluated using the inner folds of the 10-fold nested cross-validation, and the final performance of the model was measured as an average score (of Pearson's correlation and R^2 score for regression, or AUC score for classification) across the outer loop of the cross-validation (Scheinost et al., 2019; Varoquaux, 2018; Kohavi et al., 1995; Pereira et al., 2009). The model was also trained multiple times with different initialization of parameters θ in order to avoid local minima. Furthermore, it is essential to consider the trade-off between the number of parameters and the model's performance. In addition to nested cross-validation, we therefore evaluated the marginal log-likelihood (MLL), which automatically balances model fit and complexity (Rasmussen, 2006). The MLL for GPs

includes a term for model complexity based on the chosen covariance function (e.g. linear kernel, additional hyperparameters, etc.). Using the MLL for model selection is expected to favor the simplest model that can adequately explain the data, even for high-dimensional inputs such as brain scans. Although it can be challenging to compute MLL, there are exact analytical expressions for the marginal log-likelihood for GPR (with Gaussian noise as in this study) and approximation using the Laplace method for GPC (see Appendix A). All applications using GP inference and prediction on MRI data in this paper were performed using custom-made implementation in Python. The code for the GP-MKL model is available at https://github.com/neuroprognosis/GPMKL.

2.4. Application to real MRI sample

The DZNE longitudinal cognitive impairment and dementia (DELCODE) cohort is a multi-centric observational study collected at 10 sites of the German Center of Neurodegenerative Diseases (DZNE). The full sample at baseline includes 1079 participants representing a broad spectrum ranging from healthy towards clinically diagnosed as dementia. More specifically these include 236 healthy controls (HC) without any cognitive impairment, 444 subjects with subjective cognitive decline (SCD), 191 cases with mild cognitive impairment (MCI), 126 Alzheimer's patients (AD) and 82 first-degree relatives of AD patients (ADR) (Jessen et al., 2018). Subjects in the HC and ADR groups were recruited by public advertisement, whereas SCD, MCI and AD subjects through referrals (including self-referrals) in the participating memory centers. Out of 1079 subjects, 973 subjects were between 60 and 89 years old and had T1-weighted data. Of these 973 subjects, 14 subjects were excluded due to MR artifacts and poor processing quality (see details below). The inclusion criteria for the DELCODE study were as follows: SCD, MCI and AD participants were clinically assessed including medical history, psychiatric and neurological examination, neuropsychological testing, blood laboratory work-up, and routine MRI, all according to the local standards. SCD subjects were defined by the presence of subjective cognitive decline with a test performance better than −1.5 standard deviations (SD) below the age, sex, and education-adjusted normal performance on all subtests of the Consortium to Establish a Registry for Alzheimer's Disease (CERAD) neuropsychological battery. The MCI group included individuals with amnestic MCI, defined by observed cognitive decline and age, sex, and education-adjusted performance below –1.5 SD on the delayed recall trial of the CERAD word-list episodic memory tests (Jessen et al., 2014; Molinuevo et al., 2017). Moreover, subjects with mild Alzheimer's dementia had a score of at least ≥ 18 on the Mini-Mental-State Examination (MMSE) according to the recommendation from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease (McKhann et al., 2011).

Via advertisement text the HC and ADR participants were explicitly inquired to feel healthy and without relevant cognitive problems. All subjects that responded to the advertisement were screened by telephone with regard to SCD criteria (above). The report of very subtle cognitive decline, which (A) did not cause any subjective concerns; and (B) was considered normal for the age by the individual, was not an exclusion criterion for the HC group. Both the HC and SCD groups had to achieve unimpaired cognitive performance according to the same definition. Additional inclusion criteria for all groups were an age of at least 60 years, fluent German language skills, capacity to provide informed consent, and presence of a study partner. The descriptive statistics of 959 subjects are summarized in Table 1.[1]

[1] DELCODE is retrospectively registered at the German Clinical Trials Register (DRKS00007966), (04/05/2015). The study has been approved by the ethics commission and the local institutional review boards.

2.5. Memory performance assessment

The DELCODE neuropsychological battery assessment (DELCODE-NP) includes MMSE (Folstein et al., 1975), ADAS-Cog 13 (Mohs et al., 1997), FCSRT-IR (Grober et al., 2009), WMS-R Logical Memory Story A, WMS-R Digit Span (Petermann and Lepach, 2012), semantic fluency (animals) (Lezak et al., 2004), the oral form of the Symbol–Digit–Modalities Test (Thalmann et al., 2000), Boston Naming Test (Smith, 1982), Trail Making Test A and B (Reitan, 1958), Clock Drawing, and Clock Copying (Rouleau et al., 1992). In addition, the Face Name Associative Recognition Test (Polcher et al., 2017), and a Flanker task were used to assess executive control of attention (Van Dam et al., 2013). These tests were selected to have similar compatibility with ongoing studies such as ADNI (Park et al., 2012) and WRAP (Dowling et al., 2010) and to derive cognitive domain scores (includes learning and memory, language ability, executive functions and mental processing speed, working memory and visuospatial abilities) and cognitive composite score (e.g. the "Preclinical Alzheimer cognitive composite" (PACC5) (Papp et al., 2017)) for tracking cognitive decline. We calculated the preclinical Alzheimer's cognitive composite (PACC-5) as the mean of an individual's z-standardized performance (based on the cognitively unimpaired individuals) in the FCSRT Free Recall and Total Recall, the MMSE, the Wechsler Memory Scale-R (WMS-R) Logical Memory Story A Delayed Recall, the Symbol-Digit-Modalities Test, and the sum of the two category fluency tasks (animals, grocery). Factor scores or composites have been shown to improve predictive performance (Dubois et al., 2018).

In our study, a global memory performance factor with improved psychometric properties was derived using Confirmatory Factor Analysis (CFA) for baseline measures (Wolfsgruber et al., 2020). Due to availability constraints, cross-sectional baseline predictions of memory performance were based on this CFA-based memory factor while longitudinal analyses were focussed on PACC5 score.

2.6. CSF biomarker positivity

Out of 959 subjects of the DELCODE cohort, CSF biomarker ($A\beta$42/40 and pTau) characterization was available for a subset of 453 (47.23%) subjects. The CSF biomarkers were determined using commercially available kits according to vendor specifications: V-PLEX $A\beta$ Peptide Panel 1 (6E10) Kit (K15200E) and V-PLEX Human Total Tau Kit (K151LAE) (Mesoscale Diagnostics LLC, Rockville, USA), and Innotest Phospho-Tau(181P) (81581; Fujirebio Germany GmbH, Hannover, Germany). The cutoff value of the normal and abnormal concentration of $A\beta$42/40 is defined as 0.08 pg/ml (below pathological) and of pTau 73.65 pg/ml (above pathological). These cutoffs were determined on the basis of all DELCODE Baseline CSF data (n = 527) by Gaussian mixture modeling using the R package flexmix (version 2.3-15).

2.7. MRI acquisition

MRI scans were acquired in 9 out of 10 involved DZNE sites (3T Siemens scanners: 3 TIM Trio systems, 4 Verio systems, 1 Skyra and 1 Prisma system). Our main analyses were based on whole-brain T1-weighted MPRAGE (3D GRAPPA PAT 2, 1 mm3 isotropic, 256 X 256 px, 192 slices, sagittal, 5 min, TR 2500 ms, TE 4.33 ms, TI 110 ms, FA 7°). Further ROI and covariate processing was based on the additionally available FLAIR protocol (for details see Jessen et al., 2018).

2.8. MRI preprocessing and feature generation

We prepared morphometric brain features based on computational anatomy and Voxel-based Morphometry (VBM). This was achieved by combining spatial normalization of SPM (Welcome Trust Center for Human Neuroimaging, London, UK, http://www.fil.ion.ucl.ac.uk/spm) and segmentation from CAT12 (http://www.neuro.uni-jena.de/cat/,

Table 1
Demographic information for the participants from the DELCODE cohort used in this modeling study. The memory score (see Section 2.5) is transformed using min–max normalization to the unit interval. Age & education are indicated in years.

Variable	HC	SCD	MCI	AD	ADR
No. of subjects	229	388	158	109	75
Males/females	129/94	173/200	69/82	63/44	44/31
Age (Mean ±SD)	69.46 ±5.42	71.27 ±6.06	72.91 ±5.72	75.19 ±6.25	66.27 ±4.61
Memory score (Mean ±SD)	0.77 ±0.09	0.72 ±0.11	0.49 ±0.14	0.25 ±0.10	0.76 ±0.11
Education (Mean ±SD)	14.72 ±2.75	14.82 ±2.99	14.11 ±3.21	12.84 ±3.04	14.60 ±2.77

r1615, Jena, Germany) toolbox. All T1-weighted images were corrected for bias-field inhomogeneities, non-brain tissue was stripped, the GM, WM, and CSF brain tissue types were segmented using the CAT12 segmentation algorithm with partial volume estimation to account for mixed voxels with two tissue types (Tohka et al., 2004) and adaptive maximum a posteriori (AMAP) (Rajapakse et al., 1997). Finally, all scans were iteratively registered with a study-specific template space using rigid and non-linear diffeomorphic transformations to create unmodulated & modulated (Jacobian-scaled) normalized tissue segment maps (Ashburner and Friston, 2009, 2011). Additionally, previous research (Monté-Rubio et al., 2018) has shown that using unmodulated segment images can lead to improved predictive performance in various tasks. For the current study, different morphometric features were generated and compared, including modulated and unmodulated versions, under different imaging parameters such as resolution and smoothness. These input image features were later transformed into kernel matrices for each set of features.

2.9. Accounting for covariates, risk factors and nuisance variables

Covariates and nuisance variables that correlate with the imaging data might affect the predictive performance of a model by adding variability to the data (Scheinost et al., 2019; Sanderman et al., 2006). From a clinical perspective, these variables can also affect the interpretability & performance of the model trained using neuroimaging data. Therefore, it is important to account for these variables in the predictive model. The suggestion of Rao et al. (2015, 2017) in dealing with confounds with respect to predictive modeling, is to include confounds as a predictor along with the imaging features in the model. The most commonly found demographic covariates in prediction studies are age and sex (Alfaro-Almagro et al., 2021; Rao et al., 2017; Abdulkadir et al., 2014). Importantly, the latter variables have been associated with the risk of progression towards pathological ageing and AD (Lindsay et al., 2002; Riedel et al., 2016). In addition, genetic information such as ApoE4 status has been shown to increase the risk for AD and therefore might covary with memory performance and biomarker status (Wolfsgruber et al., 2014). An important nuisance variable in multi-center studies is the acquisition site. Therefore, in this predictive modeling study we considered acquisition site as predictors to account for potential covariate and confounding effects.

2.10. Application of GP-MKL framework to the DELCODE cohort

Next, we evaluated the performance of the GP-MKL framework for providing clinically relevant predictions of (Task I) memory performance and (Task II) biomarker positivity ($A\beta$42/40 and pTau) in DELCODE study participants. A schematic overview of the applied framework is provided in Fig. 1 following the preprocessing pipeline described earlier.

2.10.1. Testing GP-MKL model variations

In order to select the optimal model candidate for the predictive tasks, the potential contributions of preprocessing parameters that influences the overall predictive and generalization performance was evaluated. Firstly, using MRI-based brain features (GM, WM, CSF) we studied (A) choices of filter size of smoothing (ranging from 0 to 15 mm FWHM) using modulated and unmodulated brain tissue types for various image resolutions (1, 2, 4 and 8 mm); (B) different choice of linear vs. non-linear kernels such as RBF and ARD kernel. The resulting combination of brain tissue features coupled with choice of kernel were fed into the GP-based predictive models to evaluate the overall performance. Additionally, we studied the influence of further aspects of the model and training sample. We investigated the impact of the size of the training data set (i.e. number of subjects) in order to enable useful extrapolations for future studies on similar prediction tasks. Moreover, as shown in previous machine learning applications we explored the potential influence of reducing the dimensions of data using principal component analysis (PCA).

2.10.2. Prediction task I: Memory performance

The individual memory performance was predicted using successively richer feature sets comprising (1) demographic covariates and ApoE4 genotype status, (2) multiple morphometric imaging features such as GM density, and (3) CSF biomarkers. First, we started by using commonly known informative demographic covariates (DEMO = [age, sex, education]) and derived one separate kernel for each of these features entering the GP-MKL model. This formed a baseline model that aimed to predict cognitive performance differences.

Second, we focused on memory prediction using the optimal evaluated imaging features from the model selection task. These optimal features are systematically compared (A) combinations of tissue classes (GM, WM, CSF) and image type (modulated vs. unmodulated) for memory performance prediction; (B) the influence of smoothing kernel size; and (C) image resolution for the resulting brain tissue features fed into GP-based predictive models. We applied a whole-brain mask to extract voxelwise features for each subject. The features were then fed into selected optimal kernels to predict memory performance (for more details, see Section 2.10.1 & 3.1).

Third, using the CSF-biomarkers features, we calculated kernel matrices for CSF biomarkers Aβ42/40 and pTau separately. These kernels were then further used in GPR model for prediction of individual memory performance. Notably, the biomarker-based predictive features were only used during our first task aiming at memory performance as a target variable.

Since neuropsychological testing, MRI and CSF biomarkers come at different costs for patients and researchers, we aimed at quantitative head-to-head comparisons. The above sources of information were used to predict performance differences that are expected to be partially redundant or complementary. We therefore evaluated selected feature combinations, resulting in more complex models and a higher number of kernels.

In order to assess the performance and validate the GP-MKL predictive model during nested 10-fold CV, the correlation between predicted and true memory score and prediction R^2 (cross-validation $R^2 = 1 - MSE(predicted - observed)/MSE(observed - \mu)$) score was estimated, where MSE is the mean squared error and μ is the mean of the observed variable (Scheinost et al., 2019).

2.10.3. Prediction task II: Biomarker positivity

To classify biomarker positivity ($A\beta$42/40(+ve/-ve) and pTau(+ve/-ve)) for the available data of 453 subjects, we used the same combinations of features that we used during the prediction of memory

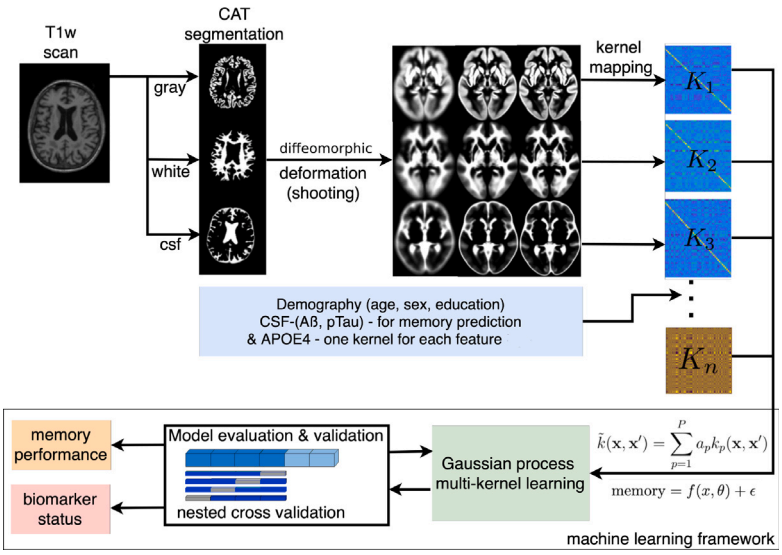

Fig. 1. GP-MKL framework. T1-weighted scans are segmented into different brain tissue types using the CAT12 segmentation algorithm. The segmented tissue types are then aligned in the same space using a non-linear image registration algorithm to obtain morphometric features. These morphometric features are mapped to kernels and forwarded to the GP-MKL framework for prediction of memory performance and biomarker status using 10 fold nested CV.

performance. To avoid double-dipping, CSF biomarker features were not used as inputs to the machine learning algorithm in this prediction task. However, we included the memory factor score obtained from neuropsychological testing in the evaluated feature combinations. For the prediction of binary biomarker positivity, the overall performance was assessed by evaluating the area under the curve (AUC) of the Receiver Operating Characteristic (ROC). Before evaluating the classifier's performance, the training and test samples were balanced across the outer and inner fold of the nested cross-validation.

2.10.4. Validation using longitudinal data

Using a predictive approach, the model explored baseline associations of brain and clinical outcomes. However, the framework has potential for clinical translational applications by predicting clinical outcome measures at baseline and follow-up measurements, i.e. prediction of future memory scores. In this final experiment, we used available longitudinal data of the DELCODE cohort to predict individual memory performance at follow-ups using baseline MRI and other covariates. Due to longitudinal availability, we focussed on the cognitive composite score PACC5 (rather than the memory factor score). We re-trained the model using baseline data (DEMO, MRI and CSF biomarkers) to predict PACC5 scores at annual follow-ups. The PACC5 score was available at baseline and five annual follow-ups for 877/695/502/373/197/41 subjects, respectively.

3. Results

3.1. Study variations of MRI features, sample size, kernel type, and PCA

Since feature engineering can be essential when constructing powerful ML models (Franke et al., 2010; Monté-Rubio et al., 2018) we studied choices of (A) filter size of smoothing (ranging from 0 to 15 mm FWHM); (B) using modulated or unmodulated brain tissue segments; (C) tissue class(es) among GM, WM, and CSF; and (D) various image resolutions (1, 2, 4 and 8 mm). Based on smoothed images of varying filter sizes, we derived separate linear kernels for each image type and brain tissue class which were subsequently entered as a linear combination in the predictive modeling framework. The effects of various combinations of input features and parameters on the accuracy of MRI-based memory prediction (i.e. only using MRI-derived features) are summarized in Fig. 2a.

Our results indicated that the combination of GM and CSF brain tissue classes for unmodulated segments at 4 mm image resolution and 4 mm smoothness (FWHM) performed best for the memory prediction task (see Table 2) in terms of R^2 score. When aiming at CSF biomarker classification, the combination of GM and CSF unmodulated tissue types performed best at 2 mm image resolution with 8 mm smoothness for Aβ42/40 and 6 mm smoothness for pTau. More generally, the choice of smoothing filter size, morphometry feature type (modulated vs. unmodulated), and the selection of tissue classes resulted in different prediction performances for all predictive tasks.

MR-based predictions using unmodulated tissue segments revealed generally a higher accuracy when compared to modulated segments. When using 1 and 2 mm image resolution, the predictive accuracy for memory performance was comparable for most smoothing filter sizes but slightly lower compared to 4 mm resolution with small smoothing kernels. Predictive accuracy remained comparably unchanged for various smaller smoothing kernels but declined gradually with more smoothing. The 8 mm input resolution generally revealed the worst overall results. Further details on statistical comparisons of different parameters can be found in (Appendix C.6, C.7, C.8).

In addition, we separately evaluated different brain tissue combinations using GM+WM+CSF (MLL = −914.5), GM+WM (MLL = −912.42), WM+CSF (MLL = −913.1). The predictive accuracy for memory prediction when combining GM+CSF (r = 0.73±0.03, R^2 = 0.51 ±0.05, MLL = −911.8) was found to be slightly higher when using the combination of all tissue classes, i.e. GM+WM+CSF (all other parameters fixed to optimal values described above). However, we found the difference not to be statistically significant (p=0.43). For CSF biomarker classification, the accuracy revealed a similar pattern for Aβ42/40 (GM+WM+CSF: AUC = 0.70 ±0.12, MLL = −372.31, GM+CSF: AUC = 0.72 ±0.12, MLL = −370.01) and pTau (GM+WM+CSF: AUC = 0.79±0.13, MLL = −287.8, GM+CSF: AUC = 0.80±0.13, MLL = −287.5). These differences were also not statistically significant (p = 0.19 and p = 0.79, respectively) but following the subtle indications for improvements based on MLL we excluded the WM tissue class in further sMRI analyses.

Comparing the performance of our models using linear, RBF, and ARD kernels, we found that linear kernels significantly outperformed both RBF and ARD kernels using different feature sets including imaging data (Fig. 2b). No significant differences were found when comparing RBF and ARD kernels. These results indicate that linear kernels might be an appropriate choice for capturing brain patterns differences in our memory prediction task.

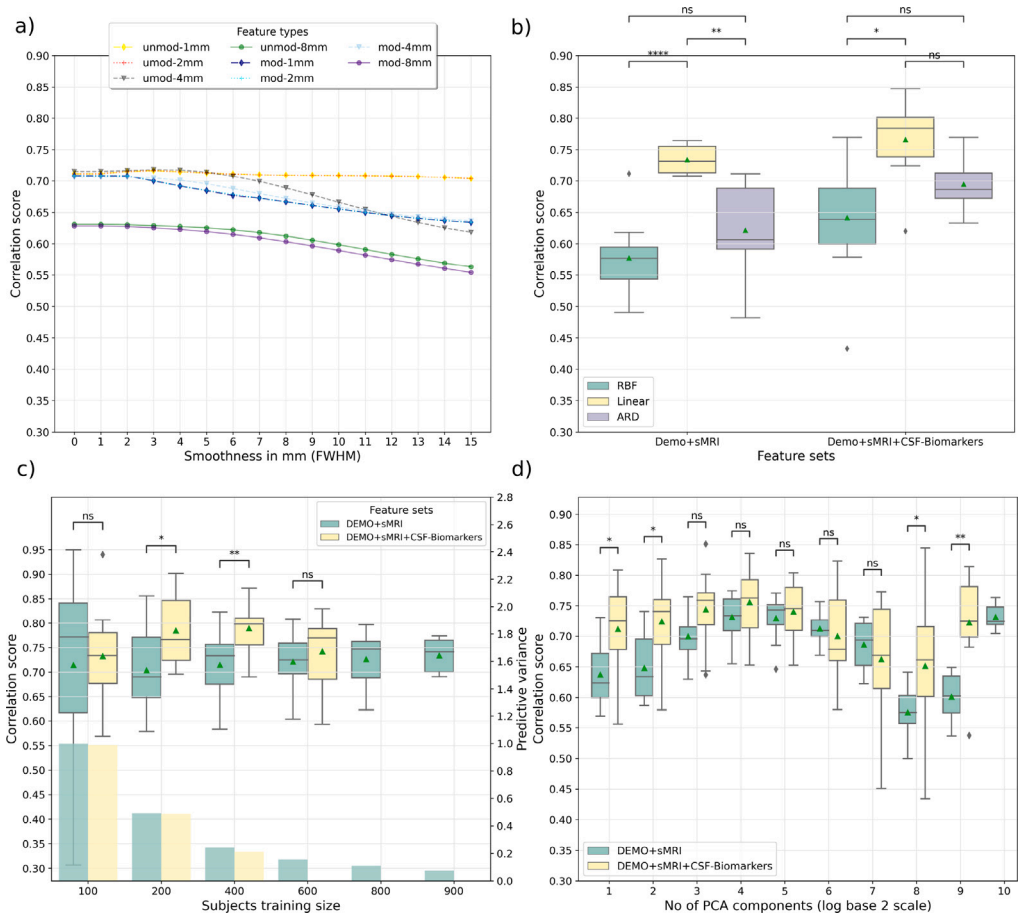

Fig. 2. Illustration of effects of choice of morphological features and kernel type on memory performance prediction using GP-MKL framework (a) Evaluating the influence of smoothing kernel, image resolution, combined tissue classes (GM+WM+CSF) and image type (modulated vs. unmodulated). Only MRI ; (b) Evaluating the predictive accuracy when using linear, squared exponential kernel (RBF) and ARD kernel. (c) Impact of training sample size; (c-yyaxis) Effect of predictive variance w.r.t sample size; and (d) Influence of reducing the dimensionality of data using Principal Component Analysis (PCA), numbers of components in $\log_2$ scale. * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$; **** $p < 0.0001$.

Table 2
Summary of the predictive accuracy for different combinations of features to predict individual memory performance. The predictive variance refers to the variance of the predictive distribution given the baseline features of an unseen subject that measures the uncertainty of individual predictions. The observed increase for models including CSF biomarkers is at least partially due to the reduced training set sample size (see also Fig. 2c yyaxis for the effect of predictive variance for different sample sizes). Kernels (hyperparameters) represents no. of kernels used in the model and no. of hyperparameters estimated by the model. MLL=marginal log-likelihood.

Features	Performance scores			Kernels	MLL
	Correlation score	R^2 score	Predictive variance		
DEMO+ApoE4	*0.53 ±0.09*	*0.28 ±0.09*	*0.1042*	*4 linear*	*−1412.9*
sMRI	*0.73 ±0.03*	*0.51 ±0.05*	*0.0755*	*2 linear*	*−911.8*
CSF Biomarkers	*0.56 ±0.13*	*0.32 ±0.14*	*0.2206*	*2 linear*	*−1162.8*
DEMO + ApoE4 + sMRI	*0.75 ±0.02*	*0.55 ±0.04*	*0.0754*	*6 linear*	*−853.2*
DEMO + ApoE4 + CSF Biomarkers	*0.64 ±0.08*	*0.40 ±0.1*	*0.2203*	*6 linear*	*−1057.5*
sMRI + CSF Biomarkers	*0.75 ±0.07*	*0.55 ±0.1*	*0.1564*	*4 linear*	*−848.3*
DEMO + ApoE4 + sMRI + CSF Biomarkers	*0.77 ±0.06*	*0.57 ±0.1*	*0.15645*	*8 linear*	*−837.2*

We further studied the potential influence of the available numbers of observations during training on predictive outcomes. We retrained the GP-MKL model in various training sample sizes ranging from 100 to 900 subjects using MRI combined with other feature sets (fixing other parameters to optimal values). Our findings suggested that performance slightly increased with the larger sample (Fig. 2c), except for the smallest sample, and estimates appeared to converge on consistent values when using at least 500 subjects. As expected, the nested cross-validation accuracy estimates showed less variation across folds when using larger sample sizes than smaller ones. Additionally, we explored

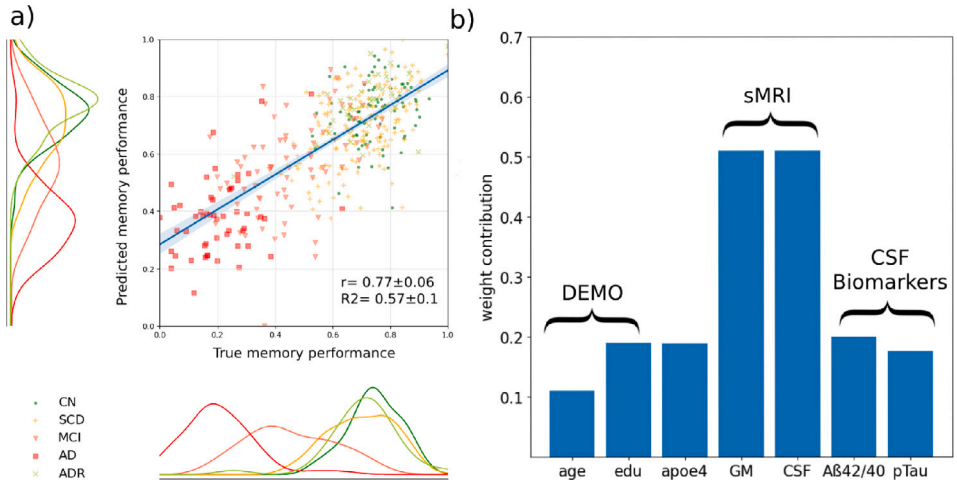

Fig. 3. (a) Illustrates prediction of individual memory performance using a combination of demographic covariates, MRI-based brain features (4 mm image resolution and 4 mm FWHM smoothness), and CSF biomarkers ($A\beta 42/40$ and pTau). (b) Weight contribution of the individual features for prediction of memory performance.

the variance of the posterior predictive distribution for different sample sizes (see Fig. 2c yy-axis) using optimal feature combinations for the memory prediction task. As expected our results indicated that the uncertainty reduced with larger training samples (see Appendix. C.8 for significance).

Next, we explored the potential benefits of reducing the dimensionality of the imaging data by applying a PCA before calculating kernels. We continued to create additional PCA components until the number of components was equal to or less than the number of subjects in the training sample. The results in Fig. 2d show that when an optimal number of PCA components is chosen a-priori or selected via cross-validation, in our case, the best performance was achieved for 32 components (r = 0.74 ±0.04, R^2 = 0.50 ±0.1. Altogether, the estimated predictive accuracy was lower than that of a non-PCA version (no dimensionality reduction) by Δr = 0.03, ΔR^2 = 0.07 (see Table 2).

Finally, we evaluated the influence of nuisance variables by including the scanning site as an additional predictor to the baseline feature set and observed a memory prediction accuracy of r = 0.53 ±0.09, R^2 = 0.28 ±0.09. The predictive accuracy remained unchanged when confounding effects of the site were incorporated as inputs to the GP-MKL models for memory prediction and therefore site was not considered in further analyses.

3.2. Prediction of memory performance

Next we aimed to assess optimal feature combinations for memory prediction. We first trained the GP-MKL framework with a simple baseline model that used only informative demographic covariates (DEMO), that is, age, sex, and education, to predict each participant's memory performance. The model performance in terms of correlation of predicted and true memory scores and predictive R^2 averaged across outer 10-fold CV was r = 0.49 ±0.11 and R^2 = 0.24 ±0.1. Additionally, including ApoE4 improved the model's performance to r = 0.53 ±0.09, R^2 = 0.28 ±0.09. Therefore, we included ApoE4 as a covariate (DEMO+ApoE4) in all further analyses.

We then explored the potential of using MRI-based brain features for improved prediction of memory performance over the above baseline model. Since the effects of variations of pre-processing and imaging features with optimal parameters were presented in the previous Section 3.1 we here focus on the summary of how demographic covariates and risk factors in combination with these 'optimal' sMRI imaging features (GM+CSF of 4 mm smoothness and image resolution) performed when predicting the participant's memory performance. A summary of out-of-sample predictive performance for all combinations of features in predicting memory is provided in Table 2.

The combination of demographics, ApoE4 and optimal sMRI features revealed a predictive accuracy of r = 0.75 ±0.02 and predictive R^2 = 0.55 ±0.04. Therefore, including specific morphometric sMRI features and demographic variables substantially enhanced overall performance over the baseline model. However, adding subject-level risk scores like age, sex, education, and ApoE4 only slightly improved the predictive performance of memory over sMRI in isolation.

Finally, we incorporated specific CSF biomarkers (available 47.23% subjects) in our predictive framework for the memory prediction task. We focused on the question of whether Aβ42/40 and pTau levels obtained from invasive lumbar puncture along with demographic and/or sMRI can further improve the prediction of each participant's memory performance.

For this purpose, we evaluated the GP-MKL model in the sub-sample with available CSF data and generated separate kernels for Aβ42/40 and pTau. When using demography, ApoE4 status, sMRI, and CSF biomarkers combined, nested cross-validation revealed accuracies of r = 0.77 ±0.06 and predictive R^2 = 0.57 ±0.1. The results of the GP-MKL model did suggest that the most important features for predicting memory performance are sMRI features, with a larger kernel weights (amplitudes) of 0.51 (shown in Fig. 3b). Further important features were found to be the cerebrospinal fluid biomarkers (Aβ42/40 and pTau), with weights of 0.2 and 0.18, respectively. Demographic features also contributed to the model's predictions, with weights of 0.12 for age, 0.19 for education, and 0.189 for ApoE4. In addition, evaluating the predictive model performance using sMRI features for the sub-sample with available CSF biomarkers yielded a comparable predictive performance as in the full sample (r = 0.72 ±0.06, R^2 = 0.50 ±0.9). The relationship between predicted and true memory performance for this model is illustrated in Fig. 3a.

In summary, the overall predictive accuracy improvement over the baseline model, which only considered covariates (e.g. from demography and ApoE4 status) increased by ΔR^2 = 0.27 for MRI and ΔR^2 = 0.12 for inclusion of Aβ42/40 and pTau biomarkers. The combination of demographic, structural brain and CSF biomarkers features revealed the highest predictive accuracies when predicting patient's memory performance. Furthermore, the uncertainty of the predictive distribution was also slightly improved for the some predictor combinations (see Table 2).

3.3. Prediction of biomarker positivity

In analogy to the above procedure, we first used the covariates (age, sex, education, and ApoE4) to predict each patient's CSF biomarker positivity. We trained our GP-MKL classifier for the DELCODE cohort

Table 3
Summary of Area Under Curve (AUC) performance scores from nested cross-validation for different combinations of features to classify CSF biomarkers positivity using the GP-MKL framework in the DELCODE cohort.

Features	Performance scores		MLL	
	AUC score Aβ42/40 (spec/sens)	AUC score pTau (spec/sens)	Aβ42/40	pTau
DEMO + ApoE4	*0.83 ±0.06 (75%/77%)*	*0.73 ±0.07 (62%/75%)*	*−228.1*	*−270.6*
sMRI	*0.72 ±0.12 (75%/54%)*	*0.80 ±0.13 (94%/34%)*	*−370.01*	*−287.5*
Memory score	*0.76 ±0.1 (78%/61%)*	*0.76 ±0.1 (71%/71%)*	*−231.1*	*−256.01*
DEMO + ApoE4 + sMRI	*0.79 ±0.12 (81%/63%)*	*0.82 ±0.13 (94%/38%)*	*−375.02*	*−305.03*
DEMO + ApoE4 + Memory	*0.86 ±0.06 (81%/76%)*	*0.78 ±0.07 (70%/74%)*	*−193.03*	*−254.03*
DEMO + ApoE4 + Memory + sMRI	*0.78 ±0.12 (79%/63%)*	*0.83 ±0.12 (95%/37%)*	*−379.04*	*−302.8*

sub-sample with available CSF biomarkers (n = 453) and assessed the classifier performance using nested cross-validation.

The accuracy for the prediction of Aβ42/40 positivity when only using these covariates was estimated with an AUC score of 0.83±0.06 (specificity = 75%, sensitivity = 77%) and for pTau, an AUC score of 0.73±0.07 (spec = 62%, sens = 75%). Similar to the memory prediction task, we explored variations of voxel-based MRI features with different tissue classes, modulation vs. no-modulation and various smoothness parameters (see Section 3.1). We then used the best-performing imaging features from the biomarker classification task and covariates to enrich the inputs.

Using the combined features from sMRI, demographics, and ApoE4, we observed a classifier performance with AUC = 0.76 ±0.12 (spec = 76%, sens = 61%) for Aβ42/40 and AUC = 0.82±0.13 (spec = 94%, sens = 38%) for pTau. The overall accuracy of pTau classification was slightly improved when using additional sMRI imaging features, while accuracy for Aβ42/40 was unexpectedly reduced.

The increase of AUC for pTau was accompanied by a marked increase in specificity to 94% while the sensitivity dropped to 38%. To investigate what is potentially driving the sensitivity differences, we trained the classification model on imaging data alone. We achieved an AUC score of 0.72±0.12 (spec = 75%, sens = 54%) for Aβ42/40 and AUC score of 0.80±0.13 (spec = 94%, sens = 34%) for pTau. The results suggested that the pattern of very high specificity and low sensitivity is instead a characteristic of the structural sMRI features during pTau classification. Table 3 provides the AUC values for selected feature combinations.

To further explore alternative feature combinations for optimal CSF biomarker classification, we additionally used the neuropsychological test-based memory performance scores to predict CSF biomarker positivity. While playing the target role in the above prediction task, memory performance is an additional input to classification algorithms predicting biomarker status. We combined kernels for memory performance with those from demographic and ApoE4 covariates (as above). We obtained an estimate of classification accuracy for Aβ42/40 of AUC = 0.86±0.06 (spec = 81%, sens = 76%) and pTau the AUC = 0.78±0.07 (spec = 70%, sens = 74%). A neuropsychological memory score provided higher predictive accuracy (specificity and sensitivity) when classifying CSF Aβ42/40 positivity compared to neuroimaging markers. For pTau, the integrated AUC of 0.82 was slightly higher when using sMRI instead of the AUC of 0.78 when using a memory test score (in combination with demographic covariates). However, the sensitivity and specificity profile was more unbalanced when using sMRI, pointing to very high specificity at the cost of sensitivity.

Finally, we combined all available baseline feature sets using neuropsychological memory performance with demographic covariates and morphometric brain features. The classifier performance was estimated as AUC = 0.78 ±0.12 (spec = 79%, sens = 63%) for Aβ42/40 and AUC = 0.83±0.12 (spec = 95%, sens = 37%) for pTau. The combination of all features only improved performance in the case of classifying pTau, while accuracy for Aβ42/40 positivity was found to be reduced compared to the above-reported combination of covariates with a neuropsychological-based memory score without MRI.

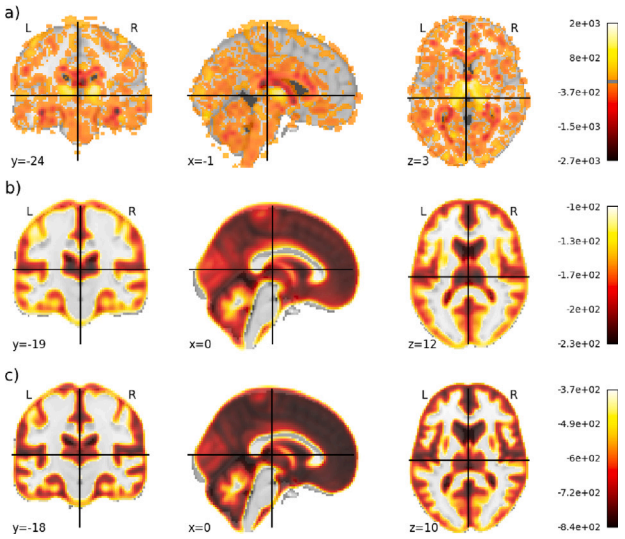

Fig. 4. (a) Illustrates the spatial weight map from sMRI features contributing to the prediction of memory performance. (b) The weight map for sMRI features during classification of *Aβ*42/40 positivity; and (c) classifying pTau positivity. Higher values on the color bar represent higher weights contribution.

Linear kernels allow visualizing weights corresponding to the contribution of each voxel during estimating predictive performance (see Appendix B for more details). Fig. 4a shows the GP-MKL weight map for predicting memory scores using the best sMRI features obtained from the regression task. The results show higher weight contribution in the brain regions of the thalamus, hippocampus, right fusiform gyrus and middle temporal gyrus. The corresponding weights for the classification of CSF biomarkers for Aβ42/40 and pTau classification are shown in Fig. 4b and c, respectively. Unexpectedly, no strong localized patterns were seen but only slightly higher weight contribution was found in the brain regions of the cerebellum, caudate nucleus, mammillary bodies and hippocampus for Aβ42/40 prediction; cerebellum, frontal gyrus and hippocampus for pTau classification.

3.4. Comparing the performance of GP-MKL and state-of-the-art model

To compare the GP-MKL model to state-of-the-art convolutional neural networks (CNNs), we used the model described by Abrol et al. (2021). The CNN model was trained on GM tissue maps with a spatial resolution of 2 mm and evaluated using 10-fold nested cross-validation. Memory performance prediction yielded a correlation value of 0.6899 ±0.0622 and an R^2 value of 0.4377 ±0.0882. Under identical conditions, the GP-MKL model achieved slightly higher correlation values of 0.6992 ±0.0262 and R^2 values of 0.4556 ±0.0350. However, the small performance advantage of the GP-MKL model is not statistically significant (p = 0.321). This suggests that both models are equally effective in analyzing GM volumes at a 2 mm spatial resolution without smoothness. Furthermore, owing to the limited sample size of CSF biomarker classification, we did not apply the CNN model to the CSF dataset.

3.5. Validation using longitudinal data

In this final experiment, we used available longitudinal data of the DELCODE cohort to predict individual memory performance scores at

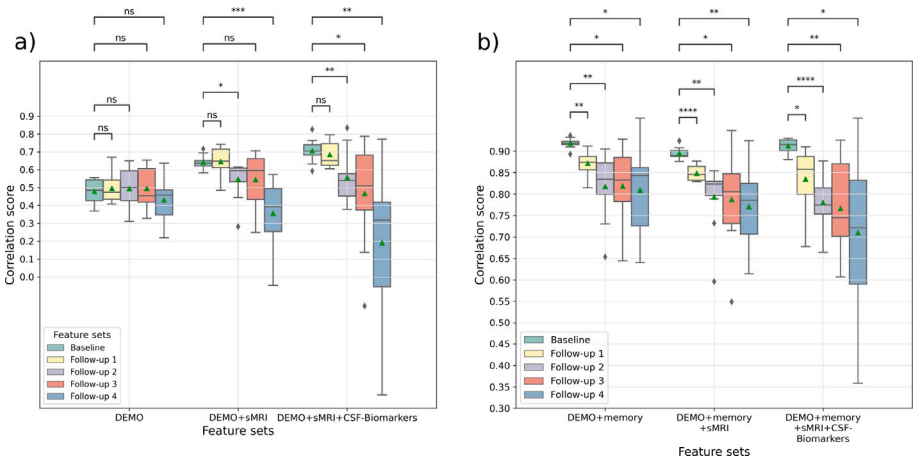

Fig. 5. Illustration of memory performance (PACC5) prediction using DEMO+ApoE4, sMRI, CSF-biomarkers for longitudinal data (a) and additionally using baseline neuropsychological memory performance (b).

annual follow-ups using baseline features including MRI. Please note that PACC5 instead of the memory factor was used during this analysis due to availability.

This model made predictions over 1, 2, 3, and 4 years and was re-trained using baseline data, including demographic covariates (age, sex, education and ApoE4), MR-based morphology, CSF Aβ42/40 and pTau to predict PACC5 scores at annual follow-ups (see methods). In this prediction over varying time gaps we obtained the highest performance when using all features for baseline (r = 0.71 ±0.07, R^2 = 0.48 ±0.09) and reduced but still decent accuracies for follow-up 1 (r = 0.69 ±0.08, R^2 = 0.43 ±0.08), follow-up 2 (r = 0.56 ±0.14, R^2 = 0.28 ±0.18), follow-up 3 (r = 0.47 ±0.3, R^2 = 0.2 ±0.3), and follow-up 4 (r = 0.2 ±0.5, R^2 = −0.08 ±0.04) respectively (Fig. 5a). One might speculate that the small size of the training sample for follow-up 4 (197 subjects) may have contributed to the result of a reduced accuracy and high variance (see also Fig. 2c). One alternative possibility is that the additional information provided by the MRI data did not outweigh the potential noise or variability introduced by including this data, e.g. due to measurement differences or minor changes to scanners. Additionally, if the characteristics of the subjects in the training sample for follow-ups are significantly different from those in the other testing samples due to selective dropout, this could have affected the model's performance. As expected, prediction accuracy significantly dropped over longer time intervals, especially after two years or more. Finally, we additionally included baseline neuropsychological memory scores as an additional predictor for performance at follow-up and observed favorable predictive performance but also minor reductions when including MRI and CSF in this comparison (Fig. 5b).

4. Discussion

In this study, we proposed a predictive model based on Gaussian processes (GPs) to evaluate the ability of various and heterogeneous information sources, namely brain morphometric features, subject-specific covariates, and CSF biomarkers, to predict individual (1) memory performance and (2) CSF-biomarker (Aβ42/40 and pTau) status. The model was tested and thoroughly evaluated on a sample of 959 individuals from the multicentric cohort (DELCODE). Our study revealed that brain morphometric features predicted memory performance the best, over and above demographics, ApoE4 status or CSF-biomarkers.

We employed a GP framework as it provides a full probabilistic prediction, which means they do provide not only the most likely prediction for a given input but also a useful measure of the uncertainty for each prediction. Moreover, GPs do offer flexibility and the ability to incorporate multiple sources of information through multi-kernel learning effectively. To explore the beneficial combinations of information from different sources, we encoded features (such as brain morphometric features, subject-specific covariates and CSF-biomarkers) as separate kernels to predict patient's memory performance and biomarker status. One key benefit of using kernel methods is the efficiency of the approach to handle high-dimensional brain images. Previous studies have demonstrated such an approximation are promising (Ashburner and Klöppel, 2011; Jollans et al., 2019), especially when combining heterogeneous multiple kernels to represent various types of predictive feature information (Zhang et al., 2012; Gupta et al., 2019; Liu et al., 2018; Gönen and Alpaydın, 2011; Wilson and Adams, 2013; Monté-Rubio et al., 2018; Dyrba et al., 2015).

4.1. The role of feature engineering in predictive modeling

A goal of this study was to identify the effective feature combinations for building predictive models using voxel-based brain morphometry. We optimized image resolution, smoothing filter size, and kernel for this purpose. First, the results showed that unmodulated morphometric features coming from gray matter and cerebrospinal fluid (GM+CSF) held similar predictive accuracy compared to those from gray matter, white matter, and cerebrospinal fluid (GM+WM+CSF). Our findings are consistent with those of Monté-Rubio et al. (2018), who found that white matter did not provide a predictive advantage over other tissue types. Interestingly, we found minor indications that unmodulated features were slightly more accurate at predicting memory and CSF biomarker status than their modulated analogs, in agreement with Monté-Rubio et al. (2018). Second, better prediction of CSF biomarker status required a higher image resolution (2 mm) than for prediction of memory performance (were no significant difference for 1, 2, and 4 mm was found), and also a stronger smoothing (6 mm for pTau, 8 mm for Aβ42/40, and 4 mm for memory). Third, our results are in line with the work of Zhu et al. (2016) showing that linear kernels outperformed ARD and RBF. Fourth, as expected for a Bayesian approach our model evaluation did suggest that if more training samples were used, the predictive performance increased and predictions were less uncertain. Fifth, dimensionality reduction, here via PCA, led to a non-significant performance drop. Taken together, our findings do suggest that feature engineering, particularly in image resolution and tuning the smoothness filter can be helpful for optimizing this GP-MKL approach and that larger study samples might further improve predictive performance and avoid the risk of overfitting.

4.2. Predicting memory performance: A comparison of predictive feature sets

We then investigated MR-based prediction of memory performance. The results demonstrated that while CSF biomarkers were better than demographics and ApoE4 status in predicting memory performance, sMRI features alone predicted them the best. Previous studies have shown that structural MRI can effectively predict cognitive scores (Zhang et al., 2012; Stonnington et al., 2010; Zhang et al., 2011; Wang et al., 2010; Duchesne et al., 2009; Liu et al., 2018). The method proposed by Zhang et al. (2012) combined features from MRI, PET, and CSF and achieved a correlation score of 0.697±0.022 and 0.739±0.012 for MMSE & ADAS-Cog scores, respectively. In our study, we aimed to predict memory performance using sMRI features and achieved a correlation score of 0.73±0.03 (Table 2). For that purpose, we used a composite score derived from multiple memory-related neuropsychological tests, which is expected to be more reliable and less noisy than individual memory test scores (such as the MMSE) (Wolfsgruber et al., 2020). Although the MMSE is a commonly used cognitive assessment tool and has a significant overlap with our composite memory score with a correlation coefficient of r = 0.75, we opted for a more comprehensive and reliable measure of memory in this study. Our results demonstrated that combining the features yielded the highest baseline cognitive memory performance (sMRI: r = 0.73±0.03, R^2 = 0.51±0.05; DEMO+ApoE4+sMRI+CSF- biomarkers: r = 0.77±0.06, R^2 = 0.570±0.1). Although our study did not achieve the same level of prediction performance as Izquierdo et al., 2017), our findings contribute valuable insights into the potential of sMRI features in predicting memory performance. Moreover, we compared our GP-MKL model to a state-of-the-art convolutional neural network (CNN) model. Our comparison revealed that our GP-MKL model achieved slightly higher correlation and R^2 values, although the difference was not statistically significant. This suggests that our GP-MKL model is as effective as the CNN model in predicting memory performance.

The availability of ApoE4 status and CSF biomarkers may be limited in certain studies due to invasiveness, costs, or time required to collect such data. However, our study shows that, when CSF is lacking, sMRI based approaches might be optimized to make predictions explaining more than 50% of the variance of memory performance in a cohort ranging from healthy ageing to MCI and AD. The fact that volumetric brain patters predicted memory performance better than CSF Aβ42/40 and pTau biomarkers could be attributed to many factors. First, cerebrospinal fluid biomarkers may reflect early brain disease signs associated with the presence of specific proteins that are not always influencing cognitive outcomes (Hu et al., 2007). However, MRI may directly reflect existing macrostructural tissue loss and atrophy in hippocampal networks (Eustache et al., 2016). Second, widespread atrophy patterns are high-dimensional and might be more reliably captured than the CSF biomarkers (Aβ42/40 & pTau). Third, MRI anatomical models can also capture aspects of a brain reserve, i.e., differences in network anatomy that positively contribute to cognitive differences in the face of brain pathology (Stern et al., 2020; Bartrés-Faz and Arenaza-Urquijo, 2011).

To summarize, a combination of structural neuroimaging markers, demographic variables, genetic information, and CSF biomarkers jointly predicted memory performance the best.

4.3. Predicting CSF biomarker status: A comparison of predictive feature sets

We then used our MR-based GP models to predict Aβ42/40 and pTau181 positivity. The results showed that demographic data was the most influential factor when predicting Aβ42/40 positivity. The combination of demographic data and memory performance produced the highest predictive accuracy corroborating findings of previous studies (Tosun et al., 2016; Jansen et al., 2018; Insel et al., 2016; Buckley et al., 2019; Ko et al., 2019; Maserejian et al., 2019; Lee et al., 2018; Ba et al., 2019; Ansart et al., 2020). Our method demonstrated a slightly higher AUC, with balanced specificity and sensitivity scores, compared to previous research. In contrast, using sMRI to classify Aβ42/40 positivity could have been more accurate, as previously reported in the literature (Tosun et al., 2013, 2016; Ansart et al., 2020; Ezzati et al., 2020). Surprisingly, we found that the accuracy of predicting Aβ42/40 positivity using morphometric features decreased slightly when sMRI features were combined with demographic and memory measures. However, we observed that using sMRI alone was the most effective predictor for pTau positivity, with high specificity (fewer false positives) and low sensitivity (more false negatives). This may suggest that pTau abnormalities, which according to the cascade model of biomarkers in AD progression are successively followed by neurodegeneration (Jack et al., 2013), might result in substantial correlations with local atrophy patterns (as opposed to amyloid). The highest accuracy in classifying pTau was achieved by combining demographic, ApoE4, sMRI, and neuropsychological memory score features. Depending on the clinical context, the classifier's threshold for distinguishing between different biomarker states could be further adjusted to prioritize high specificity or high sensitivity.

4.4. Predicting long-term cognitive performance using baseline measures

In order to support validity and utility of our model we next investigated whether baseline features might be used for predicting cognitive performance scores (PACC5) at annual longitudinal follow-ups. We revealed evidence for the longitudinal applications of the presented approach. In a previous study, Wang et al. (2010) used baseline MRI to predict future decline in cognitive performance and reported a correlation score of 0.54. Zhang et al. (2012) proposed a multi-modal multi-task learning method to predict cognitive decline in two years and achieved a correlation score of 0.52. In comparison, our GP framework achieved comparable performance (r=0.54), predicting cognitive differences two years into the future.

It is interesting to note that the DEMO+sMRI+CSF-biomarkers feature set results in Fig. 5 show an increased predictive uncertainty at follow-up 4. A plausible explanation for the reduced accuracy at follow-up 4 is the relatively small training sample size of 197 participants. An alternative explanation is that the additional information provided by the MRI data did not outweigh the potential noise and variability introduced by their inclusion. Furthermore, factors such as measurement differences and minor changes in the scanner may have introduced variations that impacted model performance. Overall, our longitudinal findings of decent predictions of follow-up support the potential of this framework for further translational applications.

4.5. Examining the interpretability of brain region contributions to prediction

It has been emphasized by Grosenick et al. (2013) that interpretability and transparency in predictive models can be improved by identifying the brain regions that contribute most to the model's performance. This framework allowed us to observe the weights determining the relationship between voxel-level signals and predicted variables, such as memory performance or biomarker positivity. In the case of memory prediction, higher weights were observed in the thalamus, hippocampus, right fusiform gyrus, and middle temporal gyrus. For predictions of CSF biomarker status, widespread distributed patterns were observed but were less simple to interpret. However, as noted by Mourao-Miranda et al. (2005), it is important to be cautious when interpreting the relationship between weights and neurocognitive processes, as the weights do not directly measure the information contained in isolated brain voxels but rather their contribution to whole pattern resulting in a predictive decision (see also Marquand et al., 2010).

4.6. Limitations of the study and future directions

The current study has several limitations that should be considered when interpreting the results. One limitation is the availability of CSF biomarkers. Despite this, the inclusion of biomarkers improved the ability to predict memory scores in the smaller sample size. We believe that the availability of CSF biomarker data for all participants would further enhance model performance and increase predictive variance, as demonstrated in previous studies and our own analysis (see Fig. 2c in Section 3.1). Additionally, training the model with a larger sample size resulted in a slight improvement in overall model performance and predictive variance. Another limitation is the limited use of additional methods such as fMRI, quantitative susceptibility maps, and white matter hyperintensities, which have been shown in previous research to improve prediction performance (Zhang et al., 2012; Gupta et al., 2019; Bouwman et al., 2007; Chételat et al., 2005). Furthermore, we used Gaussian process models to make analyze follow-up data in this study. However, these models might be extended to include both within-person and between-person models to properly accommodate the nested structures for analyzing longitudinal data (Karch et al., 2020; Abi Nader et al., 2020; Challis et al., 2015; Canas et al., 2019; Su et al., 2021; Duan et al., 2018). Additionally, our study initially included the acquisition site as a predictor, which could have introduced a bias if there is a correlation between the acquisition site and disease severity. However, our analysis indicated that including the acquisition site as a predictor did not significantly impact the results. To overcome this limitation, future research might employ data harmonization techniques as recommended by Scheinost et al. (2019), Abraham et al. (2017) and Noble et al. (2017). Finally, other effective dimensionality reduction methods, such as kernel PCA, t-distributed stochastic neighbor embedding, and uniform manifold approximation and projection, may help reduce data complexity and might further improve prediction performance.

4.7. Conclusion

In summary, our approach explored well established Bayesian GP multi-kernel methods to identify the optimal set of feature combinations to predict baseline and follow-up memory performance and CSF biomarker status. Our method achieved encouraging results, comparable or even partially outperforming some previous cutting-edge methods (Zhang et al., 2012; Stonnington et al., 2010; Zhang et al., 2011; Wang et al., 2010; Duchesne et al., 2009; Tosun et al., 2016; Jansen et al., 2018; Insel et al., 2016; Buckley et al., 2019; Ko et al., 2019; Hojjati et al., 2022; Liu et al., 2021; Yu and Xu, 2022; Liu et al., 2021; Tian et al., 2022). The framework might be further developed to support clinical decision-making in translational applications, e.g. for possible treatments that introduce a clear expectation of clinical outcomes. Future studies might explore the impact of integrating other imaging modalities (including longitudinal) and selecting feature combinations to enhance the predictive limits of memory performance (future) and biomarker status.

Disclosure

I.K. had a relationship with Biogen for consultation or advisory fees.

CRediT authorship contribution statement

A. Nemali: Carried out the experiment, Wrote the manuscript. **N. Vockert:** Provided critical feedback and helped shape the research, analysis and manuscript. **D. Berron:** Provided critical feedback and helped shape the research, analysis and manuscript. **A. Maas:** Provided critical feedback and helped shape the research, analysis and manuscript. **B.H. Schott:** Provided critical feedback and helped shape the research, analysis and manuscript. **E. Incesoy:** Provided critical feedback and helped shape the research, analysis and manuscript. **M. Dyrba:** Provided critical feedback and helped shape the research, analysis and manuscript. **L. Kleineidam:** Provided critical feedback and helped shape the research, analysis and manuscript. **E. Duzel:** Wrote the manuscript,Provided critical feedback and helped shape the research, analysis and manuscript. **G. Ziegler:** Wrote the manuscript, Provided critical feedback and helped shape the research, analysis and manuscript, Supervised the project.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The authors do not have permission to share data.

Acknowledgments

The study was supported in part by the German Center for Neurodegenerative Diseases (DZNE), Study-ID: DZNE BN012

Appendix A. Gaussian process

A.1. Gaussian process regression

In general a Gaussian process (GP) describes a (prior) distribution over functions, which is fully specified by its mean m and covariance k (for a technical introduction see Rasmussen and Williams, 2006)

$$f(\mathbf{x}) \sim \mathcal{GP}\left(m(\mathbf{x}), k\left(\mathbf{x}, \mathbf{x}'\right)\right). \tag{A.1}$$

Gaussian process regression (GPR) is a non-parametric generalization of linear regression and can be described for training data $D = \left\{\left(\mathbf{x}_i, y_i\right)\right\}_{i=1}^{n}$ as

$$y_i = f(\mathbf{x}_i) + \epsilon_i, \quad \epsilon_i \sim \mathcal{N}(0, \sigma^2) \tag{A.2}$$

with subject index i, and an additive independent identically distributed Gaussian noise with variance σ^2. In this study the latent function $f(\mathbf{x})$ incorporates our knowledge about an older participant's memory score in different locations $\mathbf{x}$ of the brain morphometric and covariate feature space describing individual differences. More specifically the prior mean and covariance are functions of the input data and imply a prior distribution over latent mappings f

$$\begin{aligned} m(\mathbf{x}) &= \mathbb{E}[f(\mathbf{x})] \\ k\left(\mathbf{x}, \mathbf{x}'\right) &= \mathbb{E}\left[(f(\mathbf{x}) - m(\mathbf{x}))\left(f\left(\mathbf{x}'\right) - m\left(\mathbf{x}'\right)\right)\right]. \end{aligned} \tag{A.3}$$

The principal idea of GPR then is that we assume that the covariance of cognitive scores $f(\mathbf{x})$ is a function of the similarities of participants' brain morphometry, expressed by a specific choice of a kernel mapping k. In this particular study, we further explored the following choices for k using linear kernels, squared exponential (SE) kernels for imaging modalities, and automatic relevance determination (ARD) for further demographic covariate inputs that effectively remove the contribution of the irrelevant input dimensions with very large length scale.

$$k_{lin}\left(\mathbf{x}, \mathbf{x}'\right) = \mathbf{x} \cdot \mathbf{x}' \tag{A.4}$$

$$k_{se}\left(\mathbf{x}, \mathbf{x}'\right) = \exp\left(-\frac{\left\|\mathbf{x} - \mathbf{x}'\right\|^2}{2\ell^2}\right) \tag{A.5}$$

$$k_{ard}\left(\mathbf{x}, \mathbf{x}'\right) = \exp\left(-\frac{1}{2}\sum_{j=1}\frac{1}{\ell_j^2}\left(x_j - x_j'\right)^2\right). \tag{A.6}$$

When using SE kernels and ARD we introduce additional hyperparameters such as characteristic length scale ℓ_j of feature dimension

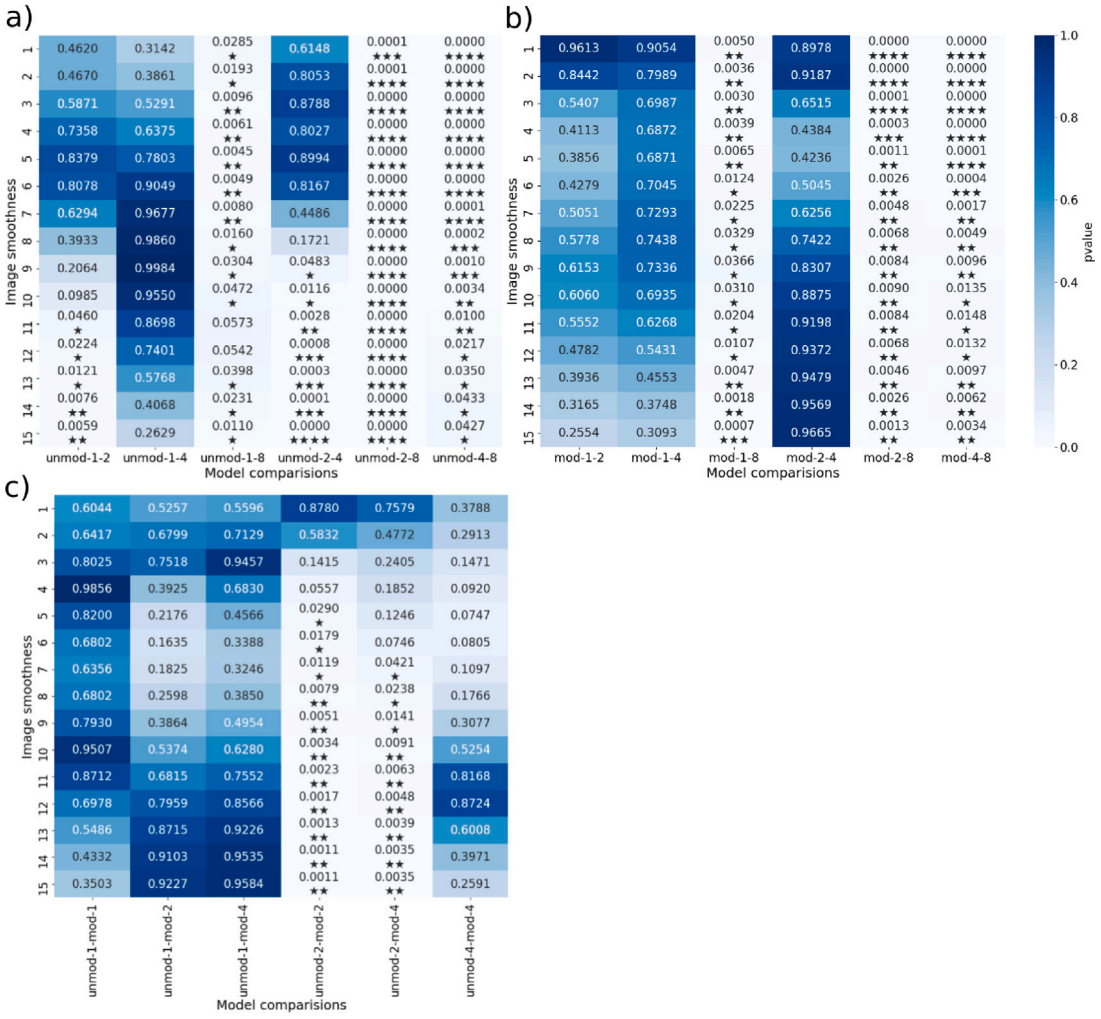

Fig. C.6. Illustrates significance of combined tissue classes (GM+WM+CSF), features (modulated vs.unmodulated) and linear kernel type on memory performance prediction using GP-MKL (see 3.1) (a) unmodulated features (b) modulated feature (c) comparisons of optimal modulated and unmodulated features * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$; **** $p < 0.0001$.

j. Having usually noisy observations the above model implies the following covariance for observed memory scores:

$$cov(\mathbf{y}) = \mathbf{K}(\mathbf{x}, \mathbf{x}') + \sigma_n^2 \mathbf{I} \tag{A.7}$$

with $\mathbf{y}$ referring to a column vector of all observed memory scores and $\mathbf{K}(\mathbf{x}, \mathbf{x}')$ denoting the evaluated kernel for all pairs of training points $\mathbf{x}$. Furthermore we are interested in predictive distributions for new test sample $\mathbf{x}^*$ given a large dataset. This can be obtained using joint distribution of training $\mathbf{y}$ and testing $\mathbf{y}^*$ samples, and then conditioning $\mathbf{y}^*|\mathbf{y}$ gives us predictive mean and variance

$$\bar{\mathbf{f}_*} = K^*(K + \sigma_n^2 I)^{-1}\mathbf{y} \tag{A.8}$$

$$\text{cov}(\mathbf{f}_*) = K^{**} - K^*(K + \sigma_n^2 I)^{-1}K^* \tag{A.9}$$

using $K = \mathbf{K}(\mathbf{x}, \mathbf{x})$, $K^* = \mathbf{K}(\mathbf{x}, \mathbf{x}^*)$ and $K^{**} = \mathbf{K}(\mathbf{x}^*, \mathbf{x}^*)$ for notation simplicity. The robustness of the GPR model is dependent on the choice of covariance function and its parameters θ. In order to choose sensible parameter estimates, the expression for the model evidence (or marginal log-likelihood) given by

$$p(\mathbf{y}|\mathbf{x}) = \int p(\mathbf{y}|f, \mathbf{x}) p(f|\mathbf{x}) df \tag{A.10}$$

using the likelihood $p(\mathbf{y}|f, \mathbf{x})$ and the prior $f|\mathbf{x} \sim \mathcal{N}(\mathbf{0}, K)$. Integrating over factorized Gaussian likelihood and prior and evaluating the terms further results in the marginal log-likelihood

$$\log p(\mathbf{y}|\mathbf{x}) = -\frac{1}{2}\mathbf{y}^\top \left(K + \sigma_n^2 I\right)^{-1} \mathbf{y} - \frac{1}{2}\log\left|K + \sigma_n^2 I\right| - \frac{n}{2}\log 2\pi \tag{A.11}$$

which is further used for model training and optimization of the (hyper-) parameters of the GP model. For efficient computation and numerical stability of matrix inversion, we use Cholesky decomposition (Rasmussen and Williams, 2006).

A.2. Gaussian process classifier

A Gaussian process classifier (GPC) for binary classification, $y_i \in \{-1, 1\}$ is a non-parametric generalization of linear logistic regression. It places a GP prior over the latent function $f(x)$ and maps it through the logistic function to model posterior class probability as $p(y = 1|\mathbf{x}) = \sigma(f(\mathbf{x}))$ (for details see Rasmussen and Williams, 2006). The inference of GPC is a two-step process. The first step is similar to GPR, which involves to compute the latent variable predictions $p(\mathbf{f}_*|\mathbf{f}) \sim \mathcal{N}(\bar{\mathbf{f}_*}, \text{cov}(\mathbf{f}_*))$ for the test sample similar to Eqs. (A.8) and (A.9). In the second step, the latent function $\mathbf{f}_*$ is squeezed through a sigmoid function to estimate class membership probability given by

$$p\left(\mathbf{y}_* = +\mathbf{1}|\mathbf{x}, \mathbf{y}, \mathbf{x}^*\right) = \int \sigma\left(\mathbf{f}_*\right) p\left(\mathbf{f}_*|\mathbf{x}, \mathbf{y}, \mathbf{x}^*\right) d\mathbf{f}_* \tag{A.12}$$

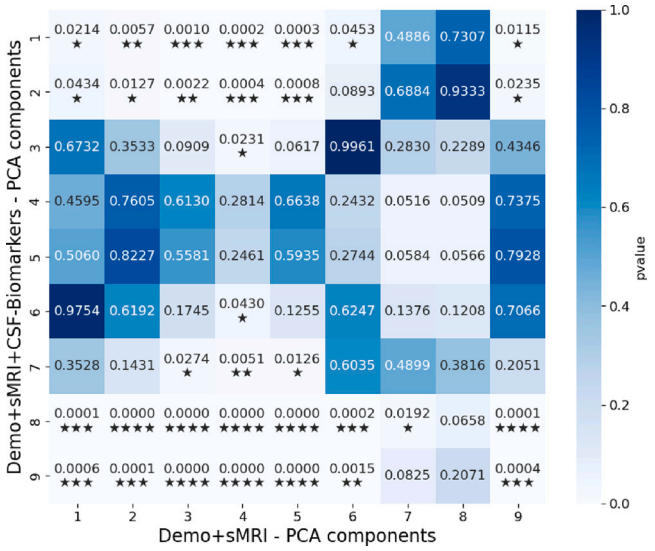

Fig. C.7. Illustrates significance to study the effect of dimensionality reduction using PCA (more details see Section 3.1 d.) * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$; **** $p < 0.0001$.

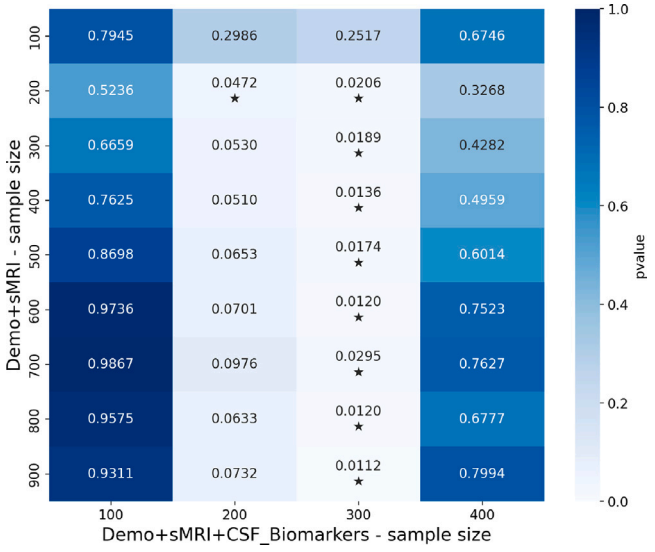

Fig. C.8. Illustrates significance studying the influence of training sample size (more details see Section 3.1 c) * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$; **** $p < 0.0001$.

accounting for all uncertainties. However in contrast to GPR Eq. (A.12) is not tractable analytically for certain sigmoid functions (such as the logistic sigmoid) and therefore we make use of the Laplace approximation to compute the approximation of the latent variable distribution

$$p(f_*|\mathbf{x},\mathbf{y},\mathbf{x}^*) = \int p(\mathbf{f}_*|\mathbf{f})p(\mathbf{f}|\mathbf{x},\mathbf{y},\mathbf{x}^*)d\mathbf{f}. \tag{A.13}$$

GPC model optimization of covariance functions and its parameters θ can be performed similarly to GPR but using an approximation of the marginal log-likelihood. More details about Laplace approximation can be found elsewhere (Bishop, 2006; Rasmussen and Williams, 2006).

Appendix B. GP weight map

The GP weights encode the contribution of voxels in predicting memory performance score and classifying biomarkers positivity. To compute the GP weight vector, we start by considering the posterior weight distribution (for more details see Rasmussen, 2006 Chapter 2) given by

$$\mathbf{w} = \sigma_n^{-2}\left(\sigma_n^{-2}\mathbf{x}^{\mathrm{T}}\mathbf{x} + \mathrm{cov}(\mathbf{x}^*)^{-1}\right)^{-1}\mathbf{x}^{\mathrm{T}}\mathbf{y} \tag{B.1}$$

Eq. (B.1) requires a huge matrix inversion, so we derivate an alternate equivalent representation, where we invert s × s matrix instead of d × d matrix (≪d, d represents number of voxels).

$$\begin{aligned}\mathbf{w} &= \mathrm{cov}(\mathbf{x}^*)\mathbf{x}^{\mathrm{T}}\left(\mathbf{x}\mathrm{cov}(\mathbf{x}^*)\mathbf{x}^{\mathrm{T}} + \sigma_n^2\mathbf{I}_m\right)^{-1}\mathbf{y} \\ &= \frac{1}{l^2}\mathbf{x}^{\mathrm{T}}\mathbf{C}^{-1}\mathbf{y} \\ &= \frac{1}{l^2}\mathbf{x}^{\mathrm{T}}\boldsymbol{\alpha}\end{aligned} \tag{B.2}$$

were $\boldsymbol{\alpha} = \mathbf{C}^{-1}\mathbf{y}$, $C = \sigma_n^2\mathbf{I}_m$. This method can be used to estimate weight contribution of each voxels for classification and regression tasks when a linear kernel is used (Schulz et al., 2017).

Appendix C. Testing for significance: Model comparisons

See Figs. C.6–C.8.

References

Abdulkadir, A., Ronneberger, O., Tabrizi, S.J., Klöppel, S., 2014. Reduction of confounding effects with voxel-wise Gaussian process regression in structural MRI. In: 2014 International Workshop on Pattern Recognition in Neuroimaging. IEEE, pp. 1–4.

Abi Nader, C., Ayache, N., Robert, P., Lorenzi, M., Initiative, A.D.N., et al., 2020. Monotonic Gaussian Process for spatio-temporal disease progression modeling in brain imaging data. Neuroimage 205, 116266.

Abraham, A., Milham, M.P., Di Martino, A., Craddock, R.C., Samaras, D., Thirion, B., Varoquaux, G., 2017. Deriving reproducible biomarkers from multi-site resting-state data: An autism-based example. NeuroImage 147, 736–745.

Abrol, A., Fu, Z., Salman, M., Silva, R., Du, Y., Plis, S., Calhoun, V., 2021. Deep learning encodes robust discriminative neuroimaging representations to outperform standard machine learning. Nat. Commun. 12 (1), 353.

Aksman, L.M., Scelsi, M.A., Marquand, A.F., Alexander, D.C., Ourselin, S., Altmann, A., ADNI, 2019. Modeling longitudinal imaging biomarkers with parametric Bayesian multi-task learning. Hum. Brain Map. 40 (13), 3982–4000.

Alfaro-Almagro, F., McCarthy, P., Afyouni, S., Andersson, J.L., Bastiani, M., Miller, K.L., Nichols, T.E., Smith, S.M., 2021. Confound modelling in UK biobank brain imaging. NeuroImage 224, 117002.

Amyot, F., Arciniegas, D.B., Brazaitis, M.P., Curley, K.C., Diaz-Arrastia, R., Gandjbakhche, A., Herscovitch, P., Hinds, S.R., Manley, G.T., Pacifico, A., et al., 2015. A review of the effectiveness of neuroimaging modalities for the detection of traumatic brain injury. J. Neurotrauma 32 (22), 1693–1721.

Ansart, M., Epelbaum, S., Gagliardi, G., Colliot, O., Dormont, D., Dubois, B., Hampel, H., Durrleman, S., Alzheimer's Disease Neuroimaging Initiative*, the INSIGHT-preAD study, 2020. Reduction of recruitment costs in preclinical AD trials: validation of automatic pre-screening algorithm for brain amyloidosis. Stat. Methods Med. Res. 29 (1), 151–164.

Arbabshirani, M.R., Plis, S., Sui, J., Calhoun, V.D., 2017. Single subject prediction of brain disorders in neuroimaging: Promises and pitfalls. Neuroimage 145, 137–165.

Ashburner, J., Friston, K.J., 2009. Computing average shaped tissue probability templates. Neuroimage 45 (2), 333–341.

Ashburner, J., Friston, K.J., 2011. Diffeomorphic registration using geodesic shooting and Gauss–Newton optimisation. NeuroImage 55 (3), 954–967.

Ashburner, J., Klöppel, S., 2011. Multivariate models of inter-subject anatomical variability. Neuroimage 56 (2), 422–439.

Ba, M., Ng, K., Gao, X., Kong, M., Guan, L., Yu, L., Alzheimer's Disease Neuroimaging Initiative, 2019. The combination of apolipoprotein E4, age and Alzheimer's disease assessment scale–cognitive subscale improves the prediction of amyloid positron emission tomography status in clinically diagnosed mild cognitive impairment. Euro. J. Neurol. 26 (5), 733–e53.

Bach, F.R., Lanckriet, G.R., Jordan, M.I., 2004. Multiple kernel learning, conic duality, and the SMO algorithm. In: Proceedings of the Twenty-First International Conference on Machine Learning. p. 6.

Barber, N., 2005. Educational and ecological correlates of IQ: A cross-national investigation. Intelligence 33 (3), 273–284.

Bartrés-Faz, D., Arenaza-Urquijo, E.M., 2011. Structural and functional imaging correlates of cognitive and brain reserve hypotheses in healthy and pathological aging. Brain Topography 24 (3), 340–357.

Beason-Held, L.L., Goh, J.O., An, Y., Kraut, M.A., O'Brien, R.J., Ferrucci, L., Resnick, S.M., 2013. Changes in brain function occur years before the onset of cognitive impairment. J. Neurosci. 33 (46), 18008–18014.

Besson, F.L., La Joie, R., Doeuvre, L., Gaubert, M., Mézenge, F., Egret, S., Landeau, B., Barré, L., Abbas, A., Ibazizene, M., et al., 2015. Cognitive and brain profiles associated with current neuroimaging biomarkers of preclinical Alzheimer's disease. J. Neurosci. 35 (29), 10402–10411.

Bishop, C.M., 2006. Pattern recognition. Mach. Learn. 128 (9).

Blennow, K., Zetterberg, H., Fagan, A.M., 2012. Fluid biomarkers in alzheimer disease. Cold Spring Harbor Perspect. Med. 2 (9), a006221.

Bouwman, F., Schoonenboom, S., van Der Flier, W., Van Elk, E., Kok, A., Barkhof, F., Blankenstein, M., Scheltens, P., 2007. CSF biomarkers and medial temporal lobe atrophy predict dementia in mild cognitive impairment. Neurobiol. Aging 28 (7), 1070–1074.

Bradley, R.H., Caldwell, B.M., 1980. The relation of home environment, cognitive competence, and IQ among males and females. Child Dev. 1140–1148.

Buckley, R.F., Sikkes, S., Villemagne, V.L., Mormino, E.C., Rabin, J.S., Burnham, S., Papp, K.V., Doré, V., Masters, C.L., Properzi, M.J., et al., 2019. Using subjective cognitive decline to identify high global amyloid in community-based samples: a cross-cohort study. Alzheimer's Dementia Diagnosis Assess. Dis. Monitoring 11 (1), 670–678.

Canas, L.S., Sudre, C.H., De Vita, E., Nihat, A., Mok, T.H., Slattery, C.F., Paterson, R.W., Foulkes, A.J., Hyare, H., Cardoso, M.J., et al., 2019. Prion disease diagnosis using subject-specific imaging biomarkers within a multi-kernel Gaussian process. NeuroImage Clin. 24, 102051.

Challis, E., Hurley, P., Serra, L., Bozzali, M., Oliver, S., Cercignani, M., 2015. Gaussian process classification of Alzheimer's disease and mild cognitive impairment from resting-state fMRI. NeuroImage 112, 232–243.

Chételat, G., Eustache, F., Viader, F., Sayette, V.D.L., Pélerin, A., Mézenge, F., Hannequin, D., Dupuy, B., Baron, J.-C., Desgranges, B., 2005. FDG-PET measurement is more accurate than neuropsychological assessments to predict global cognitive deterioration in patients with mild cognitive impairment. Neurocase 11 (1), 14–25.

Davatzikos, C., Genc, A., Xu, D., Resnick, S.M., 2001. Voxel-based morphometry using the RAVENS maps: methods and validation using simulated longitudinal atrophy. NeuroImage 14 (6), 1361–1369.

Davatzikos, C., Sotiras, A., Fan, Y., Habes, M., Erus, G., Rathore, S., Bakas, S., Chitalia, R., Gastounioti, A., Kontos, D., 2019. Precision diagnostics based on machine learning-derived imaging signatures. Magnetic Resonance Imaging 64, 49–61.

Doraiswamy, P.M., Charles, H.C., Krishnan, K.R.R., 1998. Prediction of cognitive decline in early Alzheimer's disease. Lancet 352 (9141), 1678.

Dosenbach, N.U., Nardos, B., Cohen, A.L., Fair, D.A., Power, J.D., Church, J.A., Nelson, S.M., Wig, G.S., Vogel, A.C., Lessov-Schlaggar, C.N., et al., 2010. Prediction of individual brain maturity using fMRI. Science 329 (5997), 1358–1361.

Dowling, N.M., Hermann, B., La Rue, A., Sager, M.A., 2010. Latent structure and factorial invariance of a neuropsychological test battery for the study of preclinical Alzheimer's disease. Neuropsychology 24 (6), 742.

Duan, L.L., Wang, X., Clancy, J.P., Szczesniak, R.D., 2018. Joint hierarchical Gaussian process model with application to personalized prediction in medical monitoring. Stat 7 (1), e178.

Dubois, J., Galdi, P., Paul, L.K., Adolphs, R., 2018. A distributed brain network predicts general intelligence from resting-state human neuroimaging data. Philos. Trans. R. Soc. B 373 (1756), 20170284.

Dubois, B., Hampel, H., Feldman, H.H., Scheltens, P., Aisen, P., Andrieu, S., Bakardjian, H., Benali, H., Bertram, L., Blennow, K., et al., 2016. Preclinical Alzheimer's disease: definition, natural history, and diagnostic criteria. Alzheimer's Dementia 12 (3), 292–323.

Duchesne, S., Caroli, A., Geroldi, C., Collins, D.L., Frisoni, G.B., 2009. Relating one-year cognitive change in mild cognitive impairment to baseline MRI features. Neuroimage 47 (4), 1363–1370.

Dyrba, M., Grothe, M., Kirste, T., Teipel, S.J., 2015. Multimodal analysis of functional and structural disconnection in Alzheimer's disease using multiple kernel SVM. Hum. Brain Map. 36 (6), 2118–2131.

Dyrba, M., Hanzig, M., Altenstein, S., Bader, S., Ballarini, T., Brosseron, F., Buerger, K., Cantré, D., Dechent, P., Dobisch, L., et al., 2021. Improving 3D convolutional neural network comprehensibility via interactive visualization of relevance maps: evaluation in Alzheimer's disease. Alzheimer's Res. Therapy 13 (1), 1–18.

Eustache, P., Nemmi, F., Saint-Aubert, L., Pariente, J., Péran, P., 2016. Multimodal magnetic resonance imaging in Alzheimer's disease patients at prodromal stage. J. Alzheimer's Dis. 50 (4), 1035–1050.

Ezzati, A., Harvey, D.J., Habeck, C., Golzar, A., Qureshi, I.A., Zammit, A.R., Hyun, J., Truelove-Hill, M., Hall, C.B., Davatzikos, C., et al., 2020. Predicting amyloid-β levels in amnestic mild cognitive impairment using machine learning techniques. J. Alzheimer's Dis. 73 (3), 1211–1219.

Fisher, C.K., Smith, A.M., Walsh, J.R., 2019. Machine learning for comprehensive forecasting of Alzheimer's disease progression. Sci. Rep. 9 (1), 1–14.

Folstein, M.F., Folstein, S.E., McHugh, P.R., 1975. "Mini-mental state": a practical method for grading the cognitive state of patients for the clinician. J. Psychiatric Res. 12 (3), 189–198.

Forsberg, A., Engler, H., Almkvist, O., Blomquist, G., Hagman, G., Wall, A., Ringheim, A., Långström, B., Nordberg, A., 2008. PET imaging of amyloid deposition in patients with mild cognitive impairment. Neurobiol. Aging 29 (10), 1456–1465.

Franke, K., Ziegler, G., Klöppel, S., Gaser, C., the Alzheimer Disease Neuroimaging Initiative, 2010. Estimating the age of healthy subjects from T1-weighted MRI scans using kernel methods: exploring the influence of various parameters. NeuroImage 50 (3), 883–892.

Frisoni, G.B., Fox, N.C., Jack, C.R., Scheltens, P., Thompson, P.M., 2010. The clinical use of structural MRI in Alzheimer disease. Nat. Rev. Neurol. 6 (2), 67–77.

Gaugler, J., James, B., Johnson, T., Marin, A., Weuve, J., 2019. 2019 Alzheimer's disease facts and figures. Alzheimers & Dementia 15 (3), 321–387.

Gönen, M., Alpaydın, E., 2011. Multiple kernel learning algorithms. J. Mach. Learn. Res. 12, 2211–2268.

Grober, E., Ocepek-Welikson, K., Teresi, J.A., 2009. The free and cued selective reminding test: evidence of psychometric adequacy. Psychol. Sci. Quart. 51 (3), 266–282.

Grosenick, L., Klingenberg, B., Katovich, K., Knutson, B., Taylor, J.E., 2013. Interpretable whole-brain prediction analysis with GraphNet. NeuroImage 72, 304–321.

Gupta, Y., Lama, R.K., Kwon, G.-R., Weiner, M.W., Aisen, P., Weiner, M., Petersen, R., Jack, Jr., C.R., Jagust, W., Trojanowki, J.Q., et al., 2019. Prediction and classification of alzheimer's disease based on combined features from apolipoprotein-E genotype, cerebrospinal fluid, MR, and FDG-PET imaging biomarkers. Front. Comput. Neurosci. 13, 72.

He, T., Kong, R., Holmes, A.J., Nguyen, M., Sabuncu, M.R., Eickhoff, S.B., Bzdok, D., Feng, J., Yeo, B.T., 2020. Deep neural networks and kernel regression achieve comparable accuracies for functional connectivity prediction of behavior and demographics. NeuroImage 206, 116276.

Hojjati, S.H., Babajani-Feremi, A., Initiative, A.D.N., 2022. Prediction and modeling of neuropsychological scores in Alzheimer's disease using multimodal neuroimaging data and artificial neural networks. Front. Comput. Neurosci. 15, 769982.

Hu, Y., Hosseini, A., Kauwe, J.S., Gross, J., Cairns, N.J., Goate, A.M., Fagan, A.M., Townsend, R.R., Holtzman, D.M., 2007. Identification and validation of novel CSF biomarkers for early stages of Alzheimer's disease. Proteomics–Clin. Appl. 1 (11), 1373–1384.

Humpel, C., 2011. Identifying and validating biomarkers for Alzheimer's disease. Trends Biotechnol. 29 (1), 26–32.

Insel, P.S., Palmqvist, S., Mackin, R.S., Nosheny, R.L., Hansson, O., Weiner, M.W., Mattsson, N., Alzheimer's Disease Neuroimaging Initiative, 2016. Assessing risk for preclinical β-amyloid pathology with APOE, cognitive, and demographic information. Alzheimer's & Dementia: Diagnosis Assess. Dis. Monit. 4 (1), 76–84.

Izquierdo, W., Martin, H., Cabrerizo, M., Barreto, A., Andrian, J., Rishe, N., Gonzalez-Arias, S., Loewenstein, D., Duara, R., Adjouadi, M., 2017. Robust prediction of cognitive test scores in Alzheimer's patients. In: 2017 IEEE Signal Processing in Medicine and Biology Symposium (SPMB). IEEE, pp. 1–7.

Jack, C.R., Bennett, D.A., Blennow, K., Carrillo, M.C., Feldman, H.H., Frisoni, G.B., Hampel, H., Jagust, W.J., Johnson, K.A., Knopman, D.S., et al., 2016. A/T/N: an unbiased descriptive classification scheme for Alzheimer disease biomarkers. Neurology 87 (5), 539–547.

Jack, Jr., C.R., Knopman, D.S., Jagust, W.J., Petersen, R.C., Weiner, M.W., Aisen, P.S., Shaw, L.M., Vemuri, P., Wiste, H.J., Weigand, S.D., et al., 2013. Tracking pathophysiological processes in Alzheimer's disease: an updated hypothetical model of dynamic biomarkers. Lancet Neurol. 12 (2), 207–216.

Jansen, W.J., Ossenkoppele, R., Tijms, B.M., Fagan, A.M., Hansson, O., Klunk, W.E., Van Der Flier, W.M., Villemagne, V.L., Frisoni, G.B., Fleisher, A.S., et al., 2018. Association of cerebral amyloid-β aggregation with cognitive functioning in persons without dementia. JAMA Psychiatry 75 (1), 84–95.

Jessen, F., Amariglio, R., Boxtel, M., Breteler, M., Ceccaldi, M., Chételat, G., Dubois, B., Dufouil, C., Ellis, K., Flier, W., Glodzik, L., Harten, A.V., Leon, M., McHugh, P., Mielke, M., Molinuevo, J., Mosconi, L., Osorio, R., Perrotin, A., Petersen, R., Rabin, L., Rami, L., Reisberg, B., Rentz, D., Sachdev, P., Sayette, V., Saykin, A., Scheltens, P., Shulman, M.B., Slavin, M., Sperling, R., Stewart, R., Uspenskaya, O., Vellas, B., Visser, P., Wagner, M., 2014. A conceptual framework for research on subjective cognitive decline in preclinical Alzheimer's disease. Alzheimer's Dementia 10, 844–852.

Jessen, F., Spottke, A., Boecker, H., Brosseron, F., Buerger, K., Catak, C., Fliessbach, K., Franke, C., Fuentes, M., Heneka, M.T., et al., 2018. Design and first baseline data of the DZNE multicenter observational study on predementia Alzheimer's disease (DELCODE). Alzheimer's Res. therapy 10 (1), 1–10.

Jo, T., Nho, K., Saykin, A.J., 2019. Deep learning in Alzheimer's disease: diagnostic classification and prognostic prediction using neuroimaging data. Front. Aging Neurosci. 11, 220.

Johnson, K.A., Fox, N.C., Sperling, R.A., Klunk, W.E., 2012. Brain imaging in Alzheimer disease. Cold Spring Harbor Perspect. Med. 2 (4), a006213.

Jollans, L., Boyle, R., Artiges, E., Banaschewski, T., Desrivières, S., Grigis, A., Martinot, J.-L., Paus, T., Smolka, M.N., Walter, H., et al., 2019. Quantifying performance of machine learning methods for neuroimaging data. NeuroImage 199, 351–365.

Kandiah, N., Zhang, A., Cenina, A.R., Au, W.L., Nadkarni, N., Tan, L.C., 2014. Montreal cognitive assessment for the screening and prediction of cognitive decline in early Parkinson's disease. Parkinsonism Rel. Dis. 20 (11), 1145–1148.

Karch, J.D., Brandmaier, A.M., Voelkle, M.C., 2020. Gaussian process panel modeling—machine learning inspired analysis of longitudinal panel data. Front. Psychol. 11, 351.

Knešaurek, K., 2015. Improving 18f-fluoro-d-glucose-positron emission tomography/computed tomography imaging in Alzheimer's disease studies. World J. Nucl. Med. 14 (3), 171.

Ko, H., Ihm, J.-J., Kim, H.-G., Initiative, A.D.N., et al., 2019. Cognitive profiling related to cerebral amyloid beta burden using machine learning approaches. Front. Aging Neurosci. 11, 95.

Kohavi, R., et al., 1995. A study of cross-validation and bootstrap for accuracy estimation and model selection. In: Ijcai, vol. 14 no. 2. Montreal, Canada, pp. 1137–1145.

Lee, J.H., Byun, M.S., Yi, D., Sohn, B.K., Jeon, S.Y., Lee, Y., Lee, J.-Y., Kim, Y.K., Lee, Y.-S., Lee, D.Y., 2018. Prediction of cerebral amyloid with common information obtained from memory clinic practice. Front. Aging Neurosci. 10, 309.

Lezak, M.D., Howieson, D.B., Loring, D.W., Fischer, J.S., et al., 2004. Neuropsychological assessment. Oxford University Press, USA.

Li, Z., Jiang, X., Wang, Y., Kim, Y., 2021. Applied machine learning in Alzheimer's disease research: omics, imaging, and clinical data. Emerg. topics Life Sci. 5 (6), 765–777.

Lindsay, J., Laurin, D., Verreault, R., Hébert, R., Helliwell, B., Hill, G.B., McDowell, I., 2002. Risk factors for Alzheimer's disease: a prospective analysis from the Canadian study of health and aging. Am. J. Epidemiol. 156 (5), 445–453.

Liu, J., Tian, X., Wang, J., Guo, R., Kuang, H., 2021. MTFIL-Net: automated Alzheimer's disease detection and MMSE score prediction based on feature interactive learning. In: 2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). IEEE, pp. 1002–1007.

Liu, M., Zhang, J., Adeli, E., Shen, D., 2018. Joint classification and regression via deep multi-task multi-channel learning for Alzheimer's disease diagnosis. IEEE Trans. Biomed. Eng. 66 (5), 1195–1206.

Marquand, A.F., Brammer, M., Williams, S.C., Doyle, O.M., 2014. Bayesian multi-task learning for decoding multi-subject neuroimaging data. NeuroImage 92, 298–311.

Marquand, A., Howard, M., Brammer, M., Chu, C., Coen, S., Mourão-Miranda, J., 2010. Quantitative prediction of subjective pain intensity from whole-brain fMRI data using Gaussian processes. Neuroimage 49 (3), 2178–2189.

Maserejian, N., Bian, S., Wang, W., Jaeger, J., Syrjanen, J.A., Aakre, J., Jack, Jr., C.R., Mielke, M.M., Gao, F., Initiative, A.D.N., et al., 2019. Practical algorithms for amyloid β probability in subjective or mild cognitive impairment. Alzheimer's Dementia Diagn. Assess. Dis. Monit. 11, 710–720.

Mateos-Pérez, J.M., Dadar, M., Lacalle-Aurioles, M., Iturria-Medina, Y., Zeighami, Y., Evans, A.C., 2018. Structural neuroimaging as clinical predictor: A review of machine learning applications. NeuroImage Clin. 20, 506–522.

McKhann, G.M., Knopman, D.S., Chertkow, H., Hyman, B.T., Jack, Jr., C.R., Kawas, C.H., Klunk, W.E., Koroshetz, W.J., Manly, J.J., Mayeux, R., et al., 2011. The diagnosis of dementia due to Alzheimer's disease: recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. Alzheimer's Dementia 7 (3), 263–269.

Mohs, R.C., Knopman, D., Petersen, R.C., Ferris, S.H., Ernesto, C., Grundman, M., Sano, M., Bieliauskas, L., Geldmacher, D., Clark, C., et al., 1997. Development of cognitive instruments for use in clinical trials of antidementia drugs: additions to the Alzheimer's disease assessment scale that broaden its scope. Alzheimer Dis. Assoc. Disorders.

Molinuevo, J.L., Rabin, L.A., Amariglio, R., Buckley, R., Dubois, B., Ellis, K.A., Ewers, M., Hampel, H., Klöppel, S., Rami, L., Reisberg, B., Saykin, A.J., Sikkes, S., Smart, C.M., Snitz, B.E., Sperling, R., van der Flier, W.M., Wagner, M., Jessen, F., 2017. Implementation of subjective cognitive decline criteria in research studies. Alzheimer's Dementia (ISSN: 1552-5260) 13 (3), 296–311. http://dx.doi.org/10.1016/j.jalz.2016.09.012, URL https://www.sciencedirect.com/science/article/pii/S1552526016330199.

Monté-Rubio, G.C., Falcón, C., Pomarol-Clotet, E., Ashburner, J., 2018. A comparison of various MRI feature types for characterizing whole brain anatomical differences using linear pattern recognition methods. NeuroImage 178, 753–768.

Morris, J.C., 2005. Dementia update 2005. Alzheimer Dis. Assoc. Dis. 19 (2), 100–117.

Mourao-Miranda, J., Bokde, A.L., Born, C., Hampel, H., Stetter, M., 2005. Classifying brain states and determining the discriminating activation patterns: support vector machine on functional MRI data. NeuroImage 28 (4), 980–995.

Murphy, M.P., LeVine III, H., 2010. Alzheimer's disease and the amyloid-β peptide. J. Alzheimer's Dis. 19 (1), 311–323.

Noble, S., Scheinost, D., Finn, E.S., Shen, X., Papademetris, X., McEwen, S.C., Bearden, C.E., Addington, J., Goodyear, B., Cadenhead, K.S., et al., 2017. Multisite reliability of MR-based functional connectivity. Neuroimage 146, 959–970.

Ossenkoppele, R., Schonhaut, D.R., Schöll, M., Lockhart, S.N., Ayakta, N., Baker, S.L., O'Neil, J.P., Janabi, M., Lazaris, A., Cantwell, A., et al., 2016. Tau PET patterns mirror clinical and neuroanatomical variability in Alzheimer's disease. Brain 139 (5), 1551–1567.

Papp, K.V., Rentz, D.M., Orlovsky, I., Sperling, R.A., Mormino, E.C., 2017. Optimizing the preclinical Alzheimer's cognitive composite with semantic processing: the PACC5. Alzheimer's & Dementia Transl. Res. Clin. Interventions 3 (4), 668–677.

Park, L.Q., Gross, A.L., McLaren, D.G., Pa, J., Johnson, J.K., Mitchell, M., Manly, J.J., 2012. Confirmatory factor analysis of the ADNI neuropsychological battery. Brain Imaging Behav. 6 (4), 528–539.

Pereira, F., Mitchell, T., Botvinick, M., 2009. Machine learning classifiers and fMRI: a tutorial overview. Neuroimage 45 (1), S199–S209.

Petermann, F., Lepach, A.C., 2012. Wechsler memory scale. Ed. in deutscher Übersetzung und Adaptation der WMS-IV von Davis Wechsler. Frankfurt a. M.: Pearson Assessment & Information GmbH.

Pettersson-Yeo, W., Benetti, S., Marquand, A.F., Joules, R., Catani, M., Williams, S.C., Allen, P., McGuire, P., Mechelli, A., 2014. An empirical comparison of different approaches for combining multimodal neuroimaging data with support vector machine. Front. Neurosci. 8, 189.

Polcher, A., Frommann, I., Koppara, A., Wolfsgruber, S., Jessen, F., Wagner, M., 2017. Face-name associative recognition deficits in subjective cognitive decline and mild cognitive impairment. J. Alzheimer's Dis. 56 (3), 1185–1196.

Porsteinsson, A., Isaacson, R., Knox, S., Sabbagh, M., Rubino, I., 2021. Diagnosis of early Alzheimer's disease: clinical practice in 2021. J. Prevent. Alzheimer's Dis. 8, 371–386.

Prestia, A., Caroli, A., Wade, S.K., Van Der Flier, W.M., Ossenkoppele, R., Van Berckel, B., Barkhof, F., Teunissen, C.E., Wall, A., Carter, S.F., et al., 2015. Prediction of AD dementia by biomarkers following the NIA-AA and IWG diagnostic criteria in MCI patients from three European memory clinics. Alzheimer's Dementia 11 (10), 1191–1201.

Rajapakse, J.C., Giedd, J.N., Rapoport, J.L., 1997. Statistical approach to segmentation of single-channel cerebral MR images. IEEE Trans. Med. Imaging 16 (2), 176–186.

Rakotomamonjy, A., Bach, F., Canu, S., Grandvalet, Y., 2008. Simplemkl. J. Mach. Learn. Res. 9, 2491–2521.

Rao, A., Monteiro, J.M., Ashburner, J., Portugal, L., Fernandes, O., De Oliveira, L., Pereira, M., Mourao-Miranda, J., 2015. A comparison of strategies for incorporating nuisance variables into predictive neuroimaging models. In: 2015 International Workshop on Pattern Recognition in NeuroImaging. IEEE, pp. 61–64.

Rao, A., Monteiro, J.M., Mourao-Miranda, J., Initiative, A.D., et al., 2017. Predictive modelling using neuroimaging data in the presence of confounds. NeuroImage 150, 23–49.

Rasmussen, C., 2006. Advances in Gaussian processes. Adv. Neural Inform. Process. ....

Rasmussen, C.E., Williams, C.K.I., 2006. Gaussian Processes for Machine Learning. MIT Press, Cambridge.

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

Reitan, R.M., 1958. Validity of the trail making test as an indicator of organic brain damage. Perceptual Motor Skills 8 (3), 271–276.

Riedel, B.C., Thompson, P.M., Brinton, R.D., 2016. Age, APOE and sex: triad of risk of Alzheimer's disease. J. Steroid Biochem. Molecular Biol. 160, 134–147.

Rohde, T.E., Thompson, L.A., 2007. Predicting academic achievement with cognitive ability. Intelligence 35 (1), 83–92.

Rouleau, I., Salmon, D.P., Butters, N., Kennedy, C., McGuire, K., 1992. Quantitative and qualitative analyses of clock drawings in Alzheimer's and Huntington's disease. Brain Cognit. 18 (1), 70–87.

Salvatore, C., Battista, P., Castiglioni, I., 2016. Frontiers for the early diagnosis of AD by means of MRI brain imaging and support vector machines. Curr. Alzheimer Res. 13 (5), 509–533.

Sanderman, R., Coyne, J.C., Ranchor, A.V., 2006. Age: Nuisance variable to be eliminated with statistical control or important concern? Patient Educ. Counsel. 61 (3), 315–316.

Scheinost, D., Noble, S., Horien, C., Greene, A.S., Lake, E.M., Salehi, M., Gao, S., Shen, X., O'Connor, D., Barron, D.S., et al., 2019. Ten simple rules for predictive modeling of individual differences in neuroimaging. NeuroImage 193, 35–45.

Schulz, E., Speekenbrink, M., Krause, A., 2017. A tutorial on Gaussian process regression: Modelling, exploring, and exploiting functions. http://dx.doi.org/10.1101/095190, bioRxiv arXiv:https://www.biorxiv.org/content/early/2017/10/10/095190.full.pdf.

Shawe-Taylor, J., Cristianini, N., 2004. Kernel methods for pattern analysis. p. 462.

Shawe-Taylor, J., Cristianini, N., et al., 2004. Kernel Methods for Pattern Analysis. Cambridge University Press.

Smith, A., 1982. Symbol digit modalities test (SDMT) manual (revised) western psychological services. Los Angeles.

Stern, Y., Arenaza-Urquijo, E.M., Bartrés-Faz, D., Belleville, S., Cantilon, M., Chetelat, G., Ewers, M., Franzmeier, N., Kempermann, G., Kremen, W.S., et al., 2020. Whitepaper: Defining and investigating cognitive reserve, brain reserve, and brain maintenance. Alzheimer's & Dementia 16 (9), 1305–1311.

Stonnington, C.M., Chu, C., Klöppel, S., Jack, Jr., C.R., Ashburner, J., Frackowiak, R.S., Initiative, A.D.N., et al., 2010. Predicting clinical scores from magnetic resonance scans in Alzheimer's disease. Neuroimage 51 (4), 1405–1413.

Su, W., Wang, X., Szczesniak, R.D., 2021. Flexible link functions in a joint hierarchical Gaussian process model. Biometrics 77 (2), 754–764.

Sui, J., Adali, T., Yu, Q., Chen, J., Calhoun, V.D., 2012. A review of multivariate methods for multimodal fusion of brain imaging data. J. Neurosci. Methods 204 (1), 68–81.

Thalmann, B., Monsch, A.U., Schneitter, M., Bernasconi, F., Aebi, C., Camachova-Davet, Z., Staehelin, H.B., 2000. The CERAD neuropsychological assessment battery (CERAD-NAB)—A minimal data set as a common tool for German-speaking Europe. Neurobiol. Aging (21), 30.

Tian, X., Liu, J., Kuang, H., Sheng, Y., Wang, J., The Alzheimer's Disease Neuroimaging Initiative, 2022. MRI-based multi-task decoupling learning for alzheimer's disease detection and MMSE score prediction: A multi-site validation. arXiv preprint arXiv:2204.01708.

Tohka, J., Zijdenbos, A., Evans, A., 2004. Fast and robust parameter estimation for statistical partial volume models in brain MRI. NeuroImage 23 (1), 84–97.

Tosun, D., Chen, Y.-F., Yu, P., Sundell, K.L., Suhy, J., Siemers, E., Schwarz, A.J., Weiner, M.W., Initiative, A.D.N., et al., 2016. Amyloid status imputed from a multimodal classifier including structural MRI distinguishes progressors from nonprogressors in a mild Alzheimer's disease clinical trial cohort. Alzheimer's & Dementia 12 (9), 977–986.

Tosun, D., Joshi, S., Weiner, M.W., Alzheimer's Disease Neuroimaging Initiative, 2013. Neuroimaging predictors of brain amyloidosis in mild cognitive impairment. Annals Neurol. 74 (2), 188–198.

Van Dam, N.T., Sano, M., Mitsis, E.M., Grossman, H.T., Gu, X., Park, Y., Hof, P.R., Fan, J., 2013. Functional neural correlates of attentional deficits in amnestic mild cognitive impairment. PLoS One 8 (1), e54035.

Varoquaux, G., 2018. Cross-validation failure: small sample sizes lead to large error bars. Neuroimage 180, 68–77.

Wang, Y., Fan, Y., Bhatt, P., Davatzikos, C., 2010. High-dimensional pattern regression using machine learning: from medical images to continuous clinical variables. Neuroimage 50 (4), 1519–1535.

Wilson, A., Adams, R., 2013. Gaussian process kernels for pattern discovery and extrapolation. In: International Conference on Machine Learning. PMLR, pp. 1067–1075.

Wolfsgruber, S., Kleineidam, L., Guski, J., Polcher, A., Frommann, I., Roeske, S., Spruth, E.J., Franke, C., Priller, J., Kilimann, I., Teipel, S., Buerger, K., Janowitz, D., Laske, C., Buchmann, M., Peters, O., Menne, F., Fuentes Casan, M., Wiltfang, J., Bartels, C., Düzel, E., Metzger, C., Glanz, W., Thelen, M., Spottke, A., Ramirez, A., Kofler, B., Fließ bach, K., Schneider, A., Heneka, M.T., Brosseron, F., Meiberth, D., Jessen, F., Wagner, M., on behalf of the DELCODE Study Group, 2020. Minor neuropsychological deficits in patients with subjective cognitive decline. Neurology (ISSN: 0028-3878) 95 (9), e1134–e1143. http://dx.doi.org/10.1212/WNL.0000000000010142, arXiv:https://n.neurology.org/content/95/9/e1134.full.pdf.

Wolfsgruber, S., Wagner, M., Schmidtke, K., Frölich, L., Kurz, A., Schulz, S., Hampel, H., Heuser, I., Peters, O., Reischies, F.M., et al., 2014. Memory concerns, memory performance and risk of dementia in patients with mild cognitive impairment. PLoS One 9 (7), e100812.

Woodard, J.L., Seidenberg, M., Nielson, K.A., Smith, J.C., Antuono, P., Durgerian, S., Guidotti, L., Zhang, Q., Butts, A., Hantke, N., et al., 2010. Prediction of cognitive decline in healthy older adults using fMRI. J. Alzheimer's Dis. 21 (3), 871–885.

Yu, W., Xu, H., 2022. Co-attentive multi-task convolutional neural network for facial expression recognition. Pattern Recognit. 123, 108401.

Zhang, D., Shen, D., Initiative, A.D.N., et al., 2012. Multi-modal multi-task learning for joint prediction of multiple regression and classification variables in Alzheimer's disease. NeuroImage 59 (2), 895–907.

Zhang, D., Wang, Y., Zhou, L., Yuan, H., Shen, D., Alzheimer's Disease Neuroimaging Initiative, et al., 2011. Multimodal classification of Alzheimer's disease and mild cognitive impairment. Neuroimage 55 (3), 856–867.

Zhu, F., Panwar, B., Dodge, H.H., Li, H., Hampstead, B.M., Albin, R.L., Paulson, H.L., Guan, Y., 2016. COMPASS: A computational model to predict changes in MMSE scores 24-months after initial assessment of Alzheimer's disease. Sci. Rep. 6 (1), 1–12.

Ziegler, G., Ridgway, G.R., Dahnke, R., Gaser, C., Alzheimer's Disease Neuroimaging Initiative, et al., 2014. Individualized Gaussian process-based prediction and detection of local and global gray matter abnormalities in elderly subjects. NeuroImage 97, 333–348.

Zu, C., Jie, B., Liu, M., Chen, S., Shen, D., Zhang, D., 2016. Label-aligned multi-task feature learning for multimodal classification of Alzheimer's disease and mild cognitive impairment. Brain Imaging Behav. 10 (4), 1148–1159.