Generating diagnostic profiles of cognitive decline and dementia using magnetoencephalography

Elliz P. Scheijbeler a,b,*, Deborah N. Schoonhoven a,b, Marjolein M.A. Engels a,b, Philip Scheltens a, Cornelis J. Stam b, Alida A. Gouw a,b,†, Arjan Hillebrand b,†

a Alzheimer Center Amsterdam, Department of Neurology, Amsterdam Neuroscience, Vrije Universiteit Amsterdam, Amsterdam UMC, Amsterdam, The Netherlands
b Department of Clinical Neurophysiology and MEG Center, Department of Neurology, Amsterdam Neuroscience, Vrije Universiteit Amsterdam, Amsterdam UMC, Amsterdam, The Netherlands

ARTICLE INFO

Article history:
Received 4 June 2021
Revised 11 October 2021
Accepted 4 November 2021
Available online 11 November 2021

Keywords:
Multi-class classification
Machine learning
Magnetoencephalography
Dementia
Cognitive decline

ABSTRACT

Accurate identification of the underlying cause(s) of cognitive decline and dementia is challenging due to significant symptomatic overlap between subtypes. This study presents a multi-class classification framework for subjects with subjective cognitive decline, mild cognitive impairment, Alzheimer’s disease, dementia with Lewy bodies, fronto-temporal dementia and cognitive decline due to psychiatric illness, trained on source-localized resting-state magnetoencephalography data. Diagnostic profiles, describing probability estimates for each of the 6 diagnoses, were assigned to individual subjects.

A balanced accuracy rate of 41% and multi-class area under the curve value of 0.75 were obtained for 6-class classification. Classification primarily depended on posterior relative delta, theta and beta power and amplitude-based functional connectivity in the beta and gamma frequency band. Dementia with Lewy bodies (sensitivity: 100%, precision: 20%) and Alzheimer’s disease subjects (sensitivity: 51%, precision: 90%) could be classified most accurately. Fronto-temporal dementia subjects (sensitivity: 11%, precision: 3%) were most frequently misclassified.

Magnetoencephalography biomarkers hold promise to increase diagnostic accuracy in a noninvasive manner. Diagnostic profiles could provide an intuitive tool to clinicians and may facilitate implementation of the classifier in the memory clinic.

© 2021 The Author(s). Published by Elsevier Inc. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/)

1. Introduction

Dementia refers to a clinical syndrome characterized by progressive cognitive decline, affecting the ability to perform everyday activities (WHO, 2020). Its clinical presentation varies greatly among individuals, with potential deficits in memory, language and/or executive domains. The global prevalence of dementia in individuals above 65 years of age is 7%, with over 60% of dementia caused by Alzheimer’s disease (AD) (López-Sanz et al., 2017). The “cognitive continuum” of AD is hypothesized to consist of (1) subjective cognitive decline (SCD), (2) mild cognitive impairment (MCI) and (3) dementia stage (Jack et al., 2018). While individuals in the former 2 groups present with cognitive symptoms that fall short of the requirements for a dementia diagnosis, both SCD and MCI elders show increased risk of progression to Alzheimer-type dementia (Slot et al., 2019; Vos et al., 2015). Degenerative causes of dementia alongside AD include dementia with Lewy bodies (DLB) and fronto-temporal dementia (FTD). Moreover, psychiatric disorders (PSY), particularly depression and schizophrenia, have been related to development of late-life cognitive decline (Onyike, 2016). The above-mentioned diagnoses make up over 80% of the Ams-
terdam Dementia Cohort, a heterogeneous and ongoing memory clinic cohort of the Amsterdam Alzheimer Centre (Van der Flier and Scheltens, 2018).

Early and accurate identification of the underlying cause(s) of cognitive decline and dementia is mandatory for defining therapeutic approaches and prognosis. The Amsterdam Alzheimer Centre employs an extensive 1-day diagnostic work-up for dementia, including medical and neurological examination, standard laboratory testing, neuropsychological testing, brain MRI, a lumbar puncture and EEG or magnetoencephalogram (MEG) (Van der Flier and Scheltens, 2018). Differential diagnostics requires significant expertise from clinicians, as current diagnostic guidelines remain relatively general and do not address the symptomatic and neuropsychological overlap between subtypes (Tong et al., 2017).

Data-driven classifiers have been proposed to provide a fast, systematic and objective way to assist clinicians during the diagnostic process. Previous classification studies have primarily focused on pairwise differential diagnostics of dementia subtypes (e.g., AD vs. DLB) (Bruun et al., 2018; Dauwan et al., 2016). While binary classifiers have yielded promising results, a real-world clinical scenario is inherently multi-class. The multi-class classification of cognitive decline and dementia subtypes is a challenging and less extensively studied problem. Tong et al. (2017) presented a 5-class classification framework for neurodegenerative disorders (i.e., SCD, AD, DLB, FTD and vascular dementia) trained on biomarkers obtained from demographic information, structural MRI and cerebrospinal fluid (CSF). Random Forest (Breiman, 2001) and RUSBoost (Seiffert et al., 2009) algorithms achieved overall accuracies of 75.2 ± 0.6% and 75.2 ± 0.8%, respectively. The results of a validation study indicated that the final model increased clinicians’ confidence in defined diagnoses, without improving overall diagnostic accuracy (Bruun et al., 2019). A lack of reliable biomarkers for dementia diagnoses other than AD may account for this stagnation (Sheikh-Bahaei et al., 2017). Improvement on accuracy of the proposed framework may be achieved by the identification and inclusion of alternative biomarkers.

Accumulating evidence suggests that M/EEG biomarkers could facilitate differentiation between dementia subtypes (Gouw and Stam, 2016; Liedorp et al., 2009; López-Sanz et al., 2017). EEG and MEG are neurophysiological techniques that provide information on neuronal activity, oscillatory dynamics and functional connectivity on a millisecond time scale. Importantly, the techniques are noninvasive and low risk. In contrast to electric fields, magnetic fields are less distorted by tissue conductivity of the scalp, skull, CSF and brain (Wolters et al., 2006). The resulting high spatial resolution of MEG allows for accurate source reconstruction (Hillebrand et al., 2005). Dementia-related MEG alterations have been reported in spectral power and functional connectivity (Engels et al., 2017; Mandal et al., 2018). Slowing of the posterior alpha rhythm has for example been reported in patients with MCI and dementia due to AD (Engels et al., 2016; Garcia et al., 2013). A shift to lower frequencies has also been demonstrated in patients with DLB using EEG (Van der Zande et al., 2018). Functional connectivity alterations in various regions and frequency bands have been linked to MCI, AD, DLB and psychiatric illness (Bokde et al., 2009; Franciotti et al., 2006; Gómez et al., 2009; Robinson and Mandell, 2014). Machine-learning based technology creates the potential to explore which of these MEG features, that are not, or not optimally quantifiable through visual inspection, could serve as reliable biomarkers for cognitive decline and dementia subtypes.

At the Amsterdam Alzheimer Center, an MEG-registration has been used as an alternative for the diagnostic EEG in an unsolicited subset of memory clinic patients since April 2015. The implementation of MEG in the diagnostic workup of memory clinic patients (i.e., MEG recording, preprocessing and reporting procedures) has previously been described by Gouw et al. (2021). The authors advocate the use of MEG outside of research settings and highlight the diagnostic value of both qualitative (i.e., visual inspection) and quantitative MEG analysis. The latter was supported by a “proof-of-concept” automated classification procedure. A random forest classifier (Breiman, 2001) trained on spectral MEG measures differentiated between subjects with SCD (n = 40) and AD (n = 40), achieving an accuracy rate of 0.84. This finding was replicated in an independent test cohort. In practice, differentiation between SCD and AD subjects is not a major challenge for clinicians. For MEG measures to be readily useful as (supportive) biomarkers, they should facilitate differentiation between dementia subtypes that are clinically similar. The current study therefore made use of a multi-class design: a classifier, trained on source-localized resting-state MEG data, differentiated between subjects with 1 of 6 memory clinic diagnoses; that is, SCD, MCI, AD, DLB, FTD or Psy. In order to provide clinicians with an indication on the level of certainty of the MEG-based classification, diagnostic profiles were introduced. These profiles describe, for individual subjects, the models’ probability estimates for all 6 diagnoses. In order to verify our findings, model performance on clinically relevant 3- and 4-class classification problems was reviewed. Finally, the most discriminative MEG biomarkers for the included diagnostic groups were identified.

2. Materials and methods

2.1. Dataset

Data were obtained from the Amsterdam Dementia Cohort (Van der Flier and Scheltens, 2018). All subjects visited the memory clinic of the Amsterdam Alzheimer Center between 2015 and 2020 and provided written informed consent for the use of their data for research purposes. The dataset included 392 subjects, of whom 329 underwent an MEG in the context of screening for dementia. Characteristics of this subcohort have previously been described by Gouw et al. (2021). 63 subjects underwent an MEG session as part of a young-onset dementia study (Engels et al., 2016). Variations in MEG recording acquisition and preprocessing pipelines will be addressed in the following sections. All subjects received a standardized diagnostic work-up including medical history taking, neurological and neuropsychological examination, laboratory testing, MRI, EEG and/or MEG and, if possible, a lumbar puncture to collect CSF (Van der Flier and Scheltens, 2018). Cognitive function was assessed through the Mini-Mental State Examination (MMSE, Folstein et al., 1975) and Montreal Cognitive Assessment (MoCA, Nasreddine et al., 2005).

Diagnoses were generated during a multidisciplinary consensus meeting, according to recent international guidelines. Subjects were diagnosed with SCD when reported cognitive complaints remained undetected through formal neuropsychological testing. MCI subjects exhibited objectively detectable cognitive deficits with preserved functional independence. Both subject groups did not meet the requirements for a diagnosis of dementia or any other neurological or psychiatric disorder (Jessen et al., 2014). A diagnosis of probable AD was assigned according to criteria of the National Institute on Aging – Alzheimer’s Association (McKhann et al., 2011). If available, CSF and/or amyloid PET scans were used to verify the presence of Alzheimer’s pathology. DLB diagnoses were based on the McKeith criteria (McKeith et al., 2005, 2017). FTD was diagnosed using the Neary Criteria (Neary et al., 1998) and revised criteria from Rascolovsky et al. (2011). Finally, psychiatric disorders were diagnosed according to the Diagnostic and Statistical Manual of Mental Disorders (American Psychiatric Association, 2013).

Demographic characteristics are presented in Table 1. MMSE (n = 388) and MoCA (n =115) scores could not be obtained for
all included subjects. There were statistically significant differences between subject groups as determined by one-way ANOVA (p < 0.05). Mann-Whitney U post hoc tests for pairwise comparison revealed that the DLB group (4%) included a significantly smaller percentage of females than all other subject groups (33%–49%). The average age at time of MEG measurement was highest in the DLB group (67.0 ± 4.2 years), followed by the MCI, AD and FTG groups (65.6 ± 6.5/63.9 ± 8.0/61.4 ± 7.7) and finally the SC and Psy groups (57.6 ± 8.9/55.0 ± 8.4). SCD subjects scored highest on the MMSE (median = 28, [20–30]) and MoCA tests (median = 26, [14–30]), whereas the lowest scores were obtained by subjects with AD (median MMSE = 22, [3–30], median MoCA = 18.5, [2–24]). Moderate (MCI/Psy) to low scores (FTD/DLB) were reported for subjects in the remaining groups. No statistically significant differences in education (Verhage, 1964) and disease duration were reported.

2.2. Data acquisition

Since April 2015, an MEG-registration has been used as an alternative for the diagnostic EEG in an unslected subset of memory clinic patients of the Amsterdam Alzheimer Center: each week, the first 2 participants on Monday morning underwent MEG recording. MEG recordings were obtained in a magnetically shielded room using a 306-channel whole-head Vectorview MEG system (Elekta Neuromag Oy, Helsinki, Finland). Subjects were in supine position. Recordings were sampled at 1250 Hz with an online anti-aliasing filter (410 Hz) and high-pass filter (0.1 Hz). A 3-digitizer (Fastrak, Polhemus, Colchester, VT, USA) was used to digitize the locations of 4 or 5 head position indicator coils, which were used to continuously record the subjects’ head position in relation to the MEG sensors. To provide an outline of the subjects’ scalp, ~500 additional points were digitized. For the young-onset dementia cohort, TI-weighted anatomical MRI scans with an in-plane resolution of 1 × 1 mm² were obtained using a 3T whole body MR scanner (Philips 3T Achieva, Best, The Netherlands) and used for co-registration using a surface-matching approach. For the memory clinic cohort, MRI templates were used for co-registration for time efficiency purposes. Very small, small, medium and large MRI templates, custom build using 77 MRIs from the Amsterdam Dementia Cohort, were fitted to the subjects’ scalp surface. The best-fitting template was used as a surrogate MRI. Previous research has reported no (systematic) bias or inconsistency for spectral and connectivity MEG metrics when using template versus native MRIs for co-registration (Douw et al., 2018).

The resting-state MEG acquisition protocol consisted of two 5-minute eyes-closed recordings. Subjects were instructed to relax but stay awake. The observed level of alpha suppression and reactivity in response to eye-opening, as well as of alpha squawk directly after eye-closure, can be indicative of abnormal brain function. Clinicians at the Amsterdam Alzheimer Center currently rely on these overt features when performing visual assessment for diagnostic purposes. Subjects who received an MEG in the context of diagnostic screening therefore opened and closed their eyes several times upon instruction during the recordings. Event markers were used to indicate exactly when subjects opened and closed their eyes. Only eyes-closed data was used for analysis. The acquisition protocol of the young-onset dementia (research) cohort included a 2-minute eyes-open condition, recorded in between the 2 eyes-closed recordings. This condition was included to prevent excessive drowsiness.

2.3. Preprocessing

The temporal extension of the Signal Space Separation (tSSS) filter (implemented in MaxFilter software, Elekta Neuromag Oy, version 2.215 (Taulu and Simola, 2006) was used to suppress correlated noise. Before estimating the tSSS coefficients, channels that contained excessive artefacts (i.e., flat, very noisy and squid-jump channels) were discarded based on visual inspection of the raw data. The (denoised) signal was then reconstructed for all sensors (Taulu et al., 2004; 2005). Source-localized activity was obtained using an atlas-based beamforming approach (Hillebrand et al., 2012). A beamformer aims to reconstruct the unique contribution that each region in the brain makes to the measured field. To do this, a spatial filter is constructed for each region that blocks contributions from all other sources (Hillebrand and Barnes, 2005; Hillebrand et al., 2005). The broad-band MEG data (0.5–70 Hz) were projected through the beamformer spatial filters in order to reconstruct time-series of neuronal activity for 90 regions-of-interest (ROIs) as included in the Automated Anatomical Labelling (AAL) atlas, that is, 78 cortical and 12 sub-cortical regions (without the cerebellum) (Gong et al., 2009; Tzourio-Mazoyer et al., 2002). To this end, the centroid voxel of each AAL region was used as representative for that ROI (Hillebrand et al., 2016). The sphere that best fitted the scalp surface obtained from the co-registered MRI scan was used as volume conductor model. The broad-band beamformer weights were computed using this volume conductor model, an equivalent current dipole (with optimum orientation determined using Singular Value Decomposition [Sekihara et al., 2004]), and the MEG data covariance matrix. Time-series of neuronal activity were obtained for each ROI by projecting sensor-level MEG data through the normalized beamformer weights, (Cheyne et al., 2007).

2.4. Time-series analysis

Visual inspection and time-series analyses were performed using in-house developed software (BrainWave, version 0.9.152.12.26, available from home.kpn.nl/stam7883/brainwave.html). Time-series were converted to ASCII format and down-sampled to 312 Hz. For
each subject, 10 nonoverlapping, artefact-free, eyes-closed epochs of 4096 samples (13.1072 s) were selected based on visual inspection. All epochs first received a quality score of 1 (= little to no eye movement, artefacts or signs of drowsiness) to 3 (= strong presence of artefacts and/or drowsiness) by an experienced assessor. The 10 epochs of highest quality were subsequently selected for each subject. All subjects had their eyes closed for the full duration of the selected epochs, in order to prevent any effect of eye-opening and closure on the classification features. Epoch length was determined based on a previous study on the stability of functional connectivity estimates at source-level (Fraschini et al., 2016).

The ROI epochs were filtered in canonical frequency bands, that is, delta (0.5–4 Hz), theta (4–8 Hz), alpha-1 (8–10 Hz), alpha-2 (10–13 Hz), beta (13–30 Hz) and gamma (30–48 Hz) using a discrete Fast Fourier Transform. The 90 ROIs of the AAL atlas were grouped into 12 subregions in order to reduce the total number of classification features (Left/Right Frontal, Central, Temporal, Parietal, Occipital and Subcortical) (Supplementary Table 1). For each epoch, relative power and functional connectivity measures were estimated in each frequency band for each of the 12 subregions as well as globally (i.e., the average over all ROIs in a subregion, or over all 90 ROIs), and averaged over epochs. This resulted in 156 classification features for each subject (12 regional relative power + 1 global relative power + 12 regional functional connectivity + 1 global functional connectivity) × 6 frequency bands).

2.4.1. Relative power

The relative power in each frequency band was computed by dividing the power in each frequency band by the total power in the 0.5–48 Hz band.

2.4.2. Functional connectivity

Functional connectivity was estimated for each epoch using the corrected Amplitude Envelope Correlation (AEC-c). The AEC metric estimates the amplitude coupling between ROIs (Brookes et al., 2012; Bruns et al., 2000). Prior to AEC estimation, the AEC-c makes use of pairwise orthogonalization to correct for spatial leakage (Brookes et al., 2012). This has been reported to be a consistent method for functional connectivity estimation (Colclough et al., 2016). Time-series of each ROI were Hilbert transformed to obtain the analytic signal. The absolute value of the analytic signal provided the power envelope. For each ROI, linear correlation coefficients between its power envelope and those of all other ROIs were computed and normalized between 0 and 1, with 0.5 indicating no functional connectivity (Breils et al., 2020). By averaging over the coefficients, a single AEC-c value was obtained for the individual ROIs (indicating the average connectivity between that ROI and the rest of the brain) and finally for the selected subregions. AEC-c values were averaged over epochs.

2.5. The Random Forest algorithm

The Random Forest (RF) algorithm (Breiman, 2001) was selected as classification model. A RF consists of a large number of bootstrap-aggregated (“bagged”) decision trees (Fig. 1). First, bootstrap samples are drawn from the original dataset and used to grow decision trees. At the top of a decision tree, this subset is “impure,” meaning it holds samples from multiple diagnostic groups. At each binary split, the observations are divided into subgroups based on a random subset of predictor variables (i.e., MEG features, $X_n$). The resulting subgroups are as different from each other as possible, while observations within each subgroup are as similar to each other as possible (i.e., impurity reduction is maximized). This is repeated until the terminal nodes of the decision tree, each representing a single diagnosis, are reached. After decision trees have been grown, prediction results can be obtained. Upon presentation of input data, each tree votes for the most likely diagnosis based on a series of conditional statements. Class probabilities (i.e., diagnostic profiles) are estimated by dividing the number of votes for each diagnosis by the number of trees in the forest. A final diagnostic label is obtained by majority voting across all decision trees.

The RF algorithm has been reported to be robust to overfitting and has important advantages over other machine learning algorithms regarding the ability to handle high-dimensional classification problems (Caruana and Niculescu-Mizil, 2006; Menze et al., 2009; Sarica et al., 2017). By making use of a variable importance measure, the RF can identify relevant features and perform a variable selection step prior to classification. The Gini impurity index is a variable importance measure based on the impurity reduction of splits (Breiman, 2001). It can be calculated for a node $n$ and number of classes $N_c$ as follows:

$$Gini(n) = 1 - \sum_{j=1}^{N_c} \left( \frac{p_j}{N} \right)^2$$

where $p_j$ is the relative frequency of class $j$ in node $n$. A split with a large decrease of impurity (i.e., a low Gini index) is considered important, and as a consequence the variables used at this split are also considered important.

2.6. Implementation details

2.6.1. Tools and settings

The RF classification models were constructed in Python (Python Software Foundation, Python Language Reference, version 3.7.3. Available at http://www.python.org) using open-source software tools from the Scikit-learn library (Pedregosa et al., 2011). A workflow diagram of the model building and evaluation process is displayed in Fig. 2. Default RF hyperparameter settings were used for classification (number of trees = 500, the number of features selected at each split $= \sqrt{\text{(number of features)}}$, and the minimum size of terminal nodes $= 1$).

2.6.2. Workflow

The dataset was split into a random training (90%) and test set (10%). The classes in the dataset were not approximately equally represented. This imbalance poses a challenge for effective classification, by causing potential “classifier bias” towards the majority class. In order to address this problem, the class distribution of the training data was rebalanced using random under-sampling (Drummond and Holte, 2003). The majority classes were down-sized to the minority class by removing observations from the dataset at random. Left-out samples were added to the test set in order to be able to utilize the full dataset. Although the main focus of this study was to prove the significance of 6-class classification results, performance on lower order classification problems was reviewed to verify our findings. Dementia subtype classification (i.e., AD, DLB, FTD) and separation of dementia subtypes from non-dementia diagnoses (i.e., AD, DLB, FTD, rest) were selected for this purpose. The random training/test set split was based on the full (6-class) dataset. The same training/test split was used for 3- and 4-class classification. The number of subjects in each class that made up the training and test sets used to perform 3-, 4- and 6-class classification are presented in Table 2. Assigned diagnostic labels, profiles and feature importance rankings were evaluated to assess model performance.
Fig. 1. Visualization of a multi-class RF. A RF consists of multiple “bagged” decision trees, each of which is grown using a random subset of samples and predictor variables ($X_n$). Upon presentation of a new instance, each tree “votes” for 1 class through a series of conditional statements (represented by the arrows). A diagnostic profile is estimated by dividing the number of votes for each diagnosis by the number of trees in the forest. The diagnostic label is obtained by majority voting across decision trees. Abbreviations: AD, Alzheimer’s disease; DLB, Dementia with Lewy bodies; MCI, Mild Cognitive Impairment.

### Table 2

Training and test sets used for 3-, 4- and 6-class classification

| Classification problem | Diagnosis          | Training set (n) | Test set (n) |
|------------------------|--------------------|------------------|--------------|
| AD vs. DLB vs. FTD     | AD                 | 24               | 103          |
|                        | DLB                | 24               | 3            |
|                        | FTD                | 24               | 9            |
|                        | Total              | 72               | 115          |
| AD vs. DLB vs. FTD vs. Rest | AD           | 24               | 103          |
|                        | DLB                | 24               | 3            |
|                        | FTD                | 24               | 9            |
|                        | Rest               | 24               | 181          |
|                        | Total              | 96               | 296          |
| SCD vs. MCI vs. AD vs. DLB vs. FTD vs. Psy | SCD        | 24               | 81           |
|                        | MCI                | 24               | 21           |
|                        | AD                 | 24               | 103          |
|                        | DLB                | 24               | 3            |
|                        | FTD                | 24               | 9            |
|                        | Psy                | 24               | 31           |
|                        | Total              | 144              | 248          |

2.7. Evaluation metrics

Classification performance at the level of diagnostic label was evaluated using the balanced accuracy, a class-wise weighted accuracy rate calculated by summing the proportion of correctly classified patients for each individual class.

**Balanced accuracy**

$$\text{Balanced accuracy} = \frac{1}{N_c} \sum_{i=1}^{N_c} \frac{\# \text{correctly classified subjects with disease } i}{\# \text{subjects with disease } i}$$ (2)

The integrated area under the receiver operating characteristic curve (Area Under the Curve, AUC) was used to assess the accuracy of the probabilistic predictions included in the diagnostic profiles. The AUC measures the likelihood that a classification model provides a higher probability score for a true diagnosis than a false diagnosis across different thresholds. The originally binary performance metric is insensitive to class imbalance and has been extended to multi-class classification problems (MAUC) by Hand and Till (2001).

Finally, sensitivity (or recall) and precision were reported to evaluate effectiveness of the model (Badillo et al., 2020). Sensitivity refers to the proportion of subjects in a diagnostic group that
was identified correctly. Mathematically, it is defined as:

\[
Sensitivity = \frac{True \, Positives}{True \, Positives + False \, Negatives}
\]  

(3)

Complementary to this, precision quantifies the proportion of positive identifications that was actually correct. It refers to the percentage of relevant positive results for each diagnostic group. The metric is defined as follows:

\[
Precision = \frac{True \, Positives}{True \, Positives + False \, Positives}
\]  

(4)

2.8. Statistical analysis

Statistical analyses to compare the demographic characteristics of the diagnostic groups were performed in SPSS (Version 25.0). One-way ANOVAs were performed to test the equality of group means, after which Mann-Whitney U post hoc tests were used to explore significant differences between group means.

To assess the statistical significance of our model, the classifier was tested against a random baseline. Label permutation tests (number of iterations = 2000) provided a distribution of chance-level balanced accuracy rates and MAUC values (Ojala and Garriga, 2010). The permutation \( p \) value represents the proportion of performance scores that were higher in the random condition than in the original condition. A \( p \) value <0.05 was considered significant.

2.9. Code and data availability

Due to privacy restrictions, the patient data that support the findings of this study are not openly available. The code used for analysis is available on request from the corresponding author.

3. Results

3.2. Classification performance

3.2.1. Diagnostic labels

Details of classification performance on the independent test set are presented in Table 3. A balanced accuracy rate of 0.41 and MAUC value of 0.75 were achieved for 6-class classification (i.e., SCD, MCI, AD, FTD, DLB, Psy). Performance rates increased with a decreasing number of classes. Highest classification accuracy was reported for 3-class classification of dementia subtypes (i.e., AD, DLB, FTD), with a balanced accuracy rate of 0.86 and MAUC value of 0.94. A balanced accuracy rate of 0.60 and MAUC value of 0.85 were achieved for 4-class classification (i.e., AD, DLB, FTD, rest). The normalized confusion matrices for 3, 4 and 6-class classification are visualized in Fig. 3. Table 2 presents the size and class distribution of the utilized training and test sets. Evaluation of chance performance, based on the null distribution obtained through permutation testing, revealed that
the MEG-based model performed significantly better than chance ($p < 0.001$).

All 3 confusion matrices indicate that subjects with DLB could best be discriminated by the RF model (100% correctly classified). Precision scores ranged between 20 and 23% (Table 4). While the model misclassified several MCI, AD and Psy subjects as DLB subjects, SCD and FTD subjects were never classified as a DLB subject.

Subjects with AD were differentiated with a sensitivity of 69%, 35% and 51% during 3-, 4- and 6-class classification, with high precision scores ($\geq 90\%$, Table 4). Only a few SCD, MCI and Psy subjects were misclassified as AD subjects. FTD subjects were most frequently misclassified during 4- and 6-class classification (11% correctly classified). Higher sensitivity for FTD (89%) than AD (68%) subjects was however reported during 3-class classification. Due to misclassified SCD, MCI, AD and Psy subjects, low precision scores (3, 11 and 27%) were obtained for FTD. Sensitivity of the 6-class classification model for differentiation between non-dementia diagnoses (i.e., SCD, MCI, Psy) ranged between 23 and 38%, with precision scores between 14 and 48%. During 4-class classification, high sensitivity (92%) and precision (66%) scores were achieved for the “rest” group that combined the 3 diagnostic groups.

### 3.2.2. Diagnostic profiles

For each subject, assignment of a diagnostic label was based on a diagnostic profile. The average diagnostic profiles of correctly classified subjects are visualized in Fig. 4. At group level, these profiles reflect the level of similarity in MEG features between diagnostic groups.

In line with the findings presented in the confusion matrices for 3-, 4- and 6-class classification (Fig. 3), subjects with DLB showed most distinct MEG features. This is reflected by the large percentage of trees (correctly) voting for the selected class (73%). MEG features of subjects with DLB showed little resemblance to diagnoses other than AD (all <5%). MEG features of AD subjects primarily revealed similarities with MCI (19%) and DLB (15%) subjects. MEG patterns of subjects with SCD, MCI, FTD and Psy showed considerable overlap, consistent with the lower sensitivity and precision scores reported for these diagnoses.

Fig. 5 illustrates the added value of diagnostic profiles and their potential for implementation in the memory clinic. Two individual diagnostic profiles of correctly classified SCD subjects (i.e., subjects with the same diagnostic label) were randomly selected as example. The subjects are presented in black and white. When a patient first visits the memory clinic of the Amsterdam Alzheimer Center, each of the 6 diagnoses has an a priori probability. These a priori probabilities, based on the 5960 patients that formed the Amsterdam Dementia Cohort in 2017 (Van der Flier and Scheltens, 2018), are presented by the block-pattern filled bars. For both subjects, MEG-based classification increased the likelihood of a SCD, FTD or Psy diagnosis. Subject A however presented more MCI and AD-like features than subject B, who primarily resembled Psy and FTD subjects. This type of within-group difference, which was observed across all diagnostic groups, would not have been captured using only diagnostic labels. Comparison of a diagnostic profile to the a priori probability distribution furthermore provides an intuitive way to evaluate model output.

#### 3.2.3. Distinctive features and regions

Not all MEG features contributed equally to the classification. The top 15 features for 6-class classification consisted of 12 relative power and 3 functional connectivity features (Fig. 6). A cut-off for visualization was set at features contributing >1% to the classification process. The complete feature ranking is presented in Supplementary Table 2. The 12 relative power features that contributed most to the classification included bilateral occipital delta, theta and beta power, bilateral parietal theta and beta power, left
temporal theta power and global beta power. The top 3 functional connectivity features were recorded from bilateral occipital regions in the beta frequency band and left parietal regions in the gamma frequency band.

4. Discussion

This study proposed a classification framework for 6 memory clinic diagnoses (i.e., SCD, MCI, AD, DLB, FTD, Psy) trained on source-localized resting-state MEG data. Regional and global relative power and functional connectivity MEG features obtained from 144 subjects were used to train RF models. Diagnostic labels and profiles (describing probability estimates for each of the 6 diagnoses) were assigned to an independent test set of 248 subjects. A balanced accuracy rate of 0.41 and MAUC value of 0.75 were obtained for 6-class classification. Subjects with DLB or AD were well differentiated by the model, with sensitivities >50%. Sensitivities for subjects with SCD, MCI and Psy ranged between 23 and 38%. Highest precision was achieved for AD subjects (90%), followed by SCD (48%), DLB (20%), MCI (16%) and Psy subjects (14%). Classification of subjects with FTD was most challenging for the model, resulting in a sensitivity of 11 and a precision score of 3%. Bilateral occipital delta, theta and beta power, bilateral parietal theta and beta power, left temporal theta power, global beta power and functional connectivity features recorded from occipital and left parietal regions in the beta and gamma frequency band contributed most to the classification. Model performance on clinically relevant 3- and 4-class classification problems was additionally reviewed in order to support these findings. Differentiation between dementia subtypes (i.e., AD, DLB, FTD) and separation of dementia subtypes from non-dementia memory clinic diagnoses (i.e., AD, DLB, FTD, rest) were selected for this purpose. Balanced accuracy rates of 0.86 and 0.60 and MAUC values of 0.94 and 0.85 were respectively achieved. Removing similar classes from the classifi-
carnation (as during 3-class classification) or reducing the number of classes by combining similar samples (as during 4-class classification) decreases the chance on classification errors, thereby increasing accuracy. A permutation test procedure, used to measure the likelihood of obtaining the observed balanced accuracy rates and MAUC values by chance, revealed that the MEG-based model performed significantly better than baseline during 3-, 4- and 6-class classification ($p < 0.001$).

Comparison of the presented findings to earlier studies is not straightforward due to variations in the utilized datasets and classification scenarios. Using a random forest model trained only on spectral MEG measures, Gouw et al. (2021) achieved an accuracy rate of 0.84 for the 2-class classification of SCD and AD subjects. Other binary classification studies have similarly demonstrated that MEG has discriminative power for dementia subtypes (Bruun et al., 2018; Dauwan et al., 2016). The strength of our approach relative to most previous dementia classification studies is that the model performed multi-class classification. Multi-class results are more similar to a real-world clinical scenario and are therefore more informative to clinicians. A previous multi-class study by (Koikkalainen et al., 2016) reported a balanced accuracy rate of 51.5% for the 4-class classification of AD, DLB, FTD and vascular dementia subjects based on structural MRI data. Tong et al. (2017) included an additional SCD group and reported a balanced accuracy rate of 75.2%, using a classification model trained on demographic, structural MRI and CSF biomarker data. Both studies reported low sensitivity for subjects with DLB (i.e., 32% and 31%). The MEG-based model presented in this study achieved a sensitivity of 100% for subjects with DLB, suggesting a valuable role for MEG in differential diagnostics of dementia. The DLB group was the smallest diagnostic group ($n = 27$) in this study, suggesting that the random under-sampling technique helped to alleviate the issue of “classifier bias” towards larger groups.

Relatively low precision scores (<20%) were achieved for MCI, DLB, FTD and Psy subjects. Unfortunately, it is difficult to obtain high sensitivity and precision at the same time. Improving precision typically reduces sensitivity and vice versa, resulting in a trade-off between a low number of false negatives (i.e., high sensitivity) and a low number of false positives (i.e., high precision). These metrics should thus always be examined together.

**Fig. 6.** Importance scores of the top 15 features for 6-class classification. The 15 most discriminative features included bilateral occipital delta power, bilateral occipital, parietal and left temporal theta power, bilateral occipital and parietal beta power, bilateral occipital beta functional connectivity and left parietal gamma functional connectivity. Global relative beta power (importance score: 0.0106) is not visualized. Abbreviation: AEC-c, corrected Amplitude Envelope Correlation.
The reported precision scores can partially be explained by the imbalanced group sizes. Even though the maximum number of true positives was obtained for the DLB group \((n = 3)\), the number of false positives in the larger AD, Psy and MCI groups \((n = 12)\) strongly influenced the precision score. The same holds true for the precision score for AD subjects. A large number of true positives \((n = 53)\) could be obtained due to the large group size, resulting in a higher precision score for this diagnostic group \((90\%)\).

Spectral MEG properties are known to be widely affected in DLB and AD. Both diagnoses have been associated with increased relative contribution of slow frequency rhythms (delta/theta) and decreased contribution of faster oscillations \((\alpha/\beta)\) over posterior regions (Berendse et al., 2000; Van der Zande et al., 2018). Visual evaluation of the MEG recordings used in this study revealed prominent slowing of the dominant posterior rhythm in DLB, and to lesser extent AD subjects (see Supplementary Figure 1). This increase in low frequency power most likely facilitated differentiation from other subject groups, resulting in effective classification of subjects with DLB (sensitivity: 100%; precision: 20%) and AD (sensitivity: 51%, precision: 90%).

In order to provide information on the location of cerebral current sources that give rise to the MEG classification features, this study performed source (rather than sensor-) level analysis. The feature importance ranking provided by the RF algorithm (Supplementary Table 2) confirms that occipital relative \(\alpha\) and \(\beta\) power and parietal relative \(\theta\) and \(\beta\) power had high discriminative power during 6-class classification. The 15 most important features for classification (Fig. 6) furthermore included functional connectivity features in the \(\beta\) and \(\gamma\) band, recorded from parietal and occipital regions. Overall, features in the \(\beta\) band had the highest discriminative value.

Similar features have been reported to be important for identifying dementia in previous M/EEG studies. Relative \(\beta\) power has for example been identified as an important feature for classification of AD and DLB subjects (Dauwan et al., 2016). The dopaminergic and severe cholinergic deficits that are associated with DLB may cause a reduction in \(\beta\) power, facilitating differentiation from AD subjects. Increased relative \(\theta\) power (López et al., 2014; Musaeus et al., 2018) and synchronization abnormalities in the \(\alpha\) and \(\beta\) band of posterior areas (parietal, temporal and occipital) have furthermore been associated with MCI and AD (Engels et al., 2017). Subjects with SCD and MCI reportedly show similar patterns of functional connectivity disruption, typically in Alzheimer-related regions (López-Sanz et al., 2017). While this supports the consideration of SCD and MCI as early stages of AD, it complicates differentiation. Connectivity features may have facilitated differentiation from subjects in the DLB, FTD and Psy groups, without accurate differentiation between the SCD, MCI and AD groups. Previous studies have reported reduced \(\alpha\) power in SCD, MCI, AD and DLB subjects (Engels et al., 2016; Lopez-Sans et al., 2017; Van der Zande et al., 2018). Although this electrophysiological change has been related to deficits in cognitive functioning, it does not appear to be disease-specific. This may explain why \(\alpha\) power features were not considered highly discriminative during 6-class classification.

The M/EEG of subjects with FTD reportedly remains normal or only mildly disturbed until the latest stage of disease (De Haan et al., 2009; Pijnenburg et al., 2008). The lack of a distinct neurophysiological correlate for FTD may have challenged differentiation from subjects in the SCD, MCI and Psy groups during 6-class classification. It also explains why a large percentage of FTD subjects \((89\%)\) was classified as a “non-dementia” subject during 4-class classification (AD vs. DLB vs. FTD vs. rest). The confusion matrix for 3-class classification (Fig. 3) however suggests that MEG may have diagnostic value for FTD subjects by allowing accurate differentiation from AD and DLB subjects. The absence of an “FTD-like” diagnostic group during this classification task could explain the higher sensitivity for FTD \((89\%)\) than AD \((69\%)\) subjects. The feature ranking of 3-class classification revealed that bilateral parietal, occipital and global \(\gamma\) power were among the most discriminative features for classification. Supplementary Table 3 displays the 28 features contributing \(\geq 1\%\) to the 3-class classification. While the \(\gamma\) band is frequently discarded from M/EEG analysis because of expected artefacts from muscle and eye movement (Hipp and Siegel, 2013; Whitham et al., 2007), this frequency band may hold information that is specific for FTD. Research on \(\gamma\) band characteristics could potentially help explain the discrepancy between the severity of clinical symptoms and the relatively normal M/EEG that is observed in FTD. Physiological interpretation of the multi-class feature rankings is however not straightforward. The ranking highlights features that helped classify individuals with a particular disorder, without necessarily capturing all, or the direction of significant electrophysiological differences between groups. Additional research is needed to elucidate how the classifier interprets the different cortical disruptions in each clinical condition.

In addition to a diagnostic label, the presented model assigned a diagnostic profile to each subject. This profile provides an indication on the level of certainty of the MEG-based diagnosis and could provide an intuitive tool to clinicians during the diagnostic process. In this study, a priori diagnostic probabilities were defined as the proportion of diagnosed cases in the Amsterdam Dementia Cohort (Van der Flier and Scheltens, 2018). Combined, these probabilities form a baseline diagnostic profile for memory clinic patients (Fig. 5). Comparison of an individual MEG-based profile to the baseline profile facilitates interpretation of model output. This approach may help narrow a differential diagnosis based on clinical and cognitive/neuropsychological assessment. If the likelihood of a single diagnosis is strongly increased in comparison to its a priori probability, this may reduce the need for collection of multiple types of invasive biomarker data.

Diagnostic profiles may also be useful for the identification of phenotypes within diagnostic groups. This could help define optimal targets for treatment selection, personalized intervention and prognosis. Evaluation of individual profiles revealed considerable variation between subjects within the same diagnostic group. Fig. 5 provides an example of this. In some subjects, the final diagnosis won majority voting across trees by a large margin, while for other subjects less diagnostic certainty could be achieved. As cognitive decline or dementia (of any cause) progresses, 2 or more symptom patterns and/or pathologies commonly coexist in the same patient (Elahi and Miller, 2017). Diagnostic profiles may reflect the accumulation of concurrent pathophysiological processes within the brain. Moreover, increasing evidence suggests that those who will progress to dementia already show M/EEG alterations specific to a subtype at an early stage. Multiple M/EEG biomarkers, mainly related to activity in the \(\beta\) and \(\theta\) band, are reportedly able to predict conversion from MCI to AD (Gouw et al., 2017; Musaeus et al., 2018; Poil et al., 2013). MCI subjects who converted to DLB, rather than AD, have (in retrospect) also been found to have baseline EEG abnormalities compatible with this disease (Bonanni et al., 2015; Van der Zande et al., 2020). Identification of individuals who are on the trajectory to develop AD or DLB, prior to the presentation of the characteristic phenotype and dementia, would allow for early intervention. This could potentially delay severe symptoms and/or institutionalization.

The presented framework did not include a healthy control group. Individuals without (subjective) memory complaints are not referred to a memory clinic, nor do they require a diagnosis. In the
future, we aim to develop a classifier that can predict disease trajectories based on MEG measures. This classifier should be able to differentiate between subjects who will remain cognitively normal over time and subjects who will convert to MCI or a specific dementia subtype. We believe that this type of “longitudinal” or predictive classifier is a better fit for inclusion of a healthy participant group than the current diagnostic framework.

This study is the first to have included a Psy group in a multi-class classification framework for cognitive decline and dementia. The reported sensitivity and precision for this diagnostic group (23 and 14%) could be explained by its heterogeneous character. The Psy group consisted of subjects with varying diagnoses (e.g., major depressive disorder, bipolar disorder) and included the largest number of subjects that used medication that can affect the recorded MEG signals (Supplementary Table 4). As we aimed to identify biomarkers that can be used in daily practice, it was essential that the model was able to deal with data of suboptimal quality. Training the model on a larger dataset might allow signal fluctuations induced by heterogeneity of symptoms and medication use to be captured better and may therefore improve classification accuracy.

4.1. Limitations

This supervised machine learning study used labelled data for model training. As described in the methods section, clinical diagnoses were based on several types of biomarker data, including visual inspection of M/EEG recordings. Although MEG was incorporated in the diagnostic work-up, the technique is not a major determinant in clinical decision making (yet) (Van der Flier and Scheltens, 2018). Potential circularity was furthermore avoided by performing classification based on MEG features that are not, or not optimally, quantifiable through visual inspection. In order to verify the reliability of the obtained diagnostic profiles, the accuracy of the model’s predicted probabilities should be quantified. Only probability scores produced by a well calibrated model can be meaningfully interpreted by clinicians. Although various performance metrics have been proposed to capture the level of model calibration (e.g., Brier (Skill) Score, Log Loss Score), a lack of consensus remains on the reliability of these metrics when applied to multi-class and imbalanced datasets (Dormann et al., 2020; Steyerberg et al., 2010). The dataset used in this study is among the largest resting-state MEG datasets presently available for dementia research. Its relatively small size however impeded the possibility to improve probability calibration through for example, Platt Transformation or Isotonic Regression (Chen et al., 2018), which require an extra cross-validation step. The best (evaluation) method to transform classifier probability scores to the probability of disease scale needs to be determined. The presented findings advocate the use of MEG in the differential and early diagnostics of dementia (in particular for DLB). It should however be noted that the number of DLB, FTD and MCI subjects in the test sets was small (Table 2). Validation on larger datasets is needed, and the presented results should therefore be interpreted with caution.

4.2. Conclusion

The presented findings demonstrate the potential for MEG biomarkers to increase diagnostic accuracy of cognitive decline and dementia in a noninvasive manner. This provides support for the implementation of MEG as a standard diagnostic tool in the memory clinic. Multi-class – rather than binary – classification was performed in order to facilitate translation of the classification model to a clinical setting. Diagnostic profiles were introduced in order to provide an indication of the level of certainty of the MEG-based diagnosis. Profiles may furthermore be helpful for a better characterization of patient groups, by reflecting potentially overlapping disease processes. Relative power features in the delta, theta and beta frequency bands combined with amplitude-based functional connectivity features in the beta and gamma band, recorded from posterior regions, had most discriminative power during 6-class classification. Subjects with DLB (sensitivity: 100%, precision: 20%) and AD (sensitivity: 51%, precision: 90%) could best be discriminated. The MEG-based diagnostic profiles of cognitive decline and dementia should be systematically compared to clinically generated differential diagnoses in order to determine the added value of this approach.

Authors’ contributions

Elliz P. Scheijbeler: Conceptualization; Data curation; Formal analysis; Visualization; Writing - original draft. Deborah N. Schoonhoven: Conceptualization; Supervision; Writing - review & editing. Marjolein M. A. Engels: Data curation; Writing - review & editing. Philip Scheltens: Writing - review & editing. Cornelis J. Stam: Conceptualization; Writing - review & editing. Alida A. Gouw: Conceptualization; Data curation; Supervision; Writing - review & editing. Arjan Hillebrand: Conceptualization; Supervision; Writing - review & editing.

Disclosure statement

The authors have no actual or potential conflicts of interest.

Acknowledgements

The authors would like to thank the participants of the Amsterdam Dementia Cohort for their contribution. Research of Amsterdam Alzheimer Center is part of the Neurodegeneration program of Amsterdam Neuroscience. The Amsterdam Alzheimer Center is supported by Alzheimer Nederland and Stichting VUmc funds. The authors thank technicians PJ, Ris, C.H. Pluige, N. Sijsma, N.C. Ake- mann, N. Zwagerman and M.C. Alting Silber for acquisition of the MEG data. This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.neurobiolaging.2021.11.002.

References

American Psychiatric Association, 2013. Diagnostic and Statistical Manual of Mental Disorders (DSM-5®). American Psychiatric Pub.
Badillo, S, Banfai, B, Birzele, F, Davydov, II, Hutchinson, L, Kam-Thong, T, Siebourg-Polster, J, Steiert, B, 2020. An introduction to machine learning. Clin. Pharmacol. Ther. 107 (4), 871–885.
Bereznude, JHW, Verbunt, JPA, Scheltens, P, Van Dijk, BW, Jonkman, EJ, 2000. Magnetoencephalographic analysis of cortical activity in AD: a pilot study. Clin. Neurophysiol. 111 (4), 604–612.
Bokde, AL, Evers, M, Hampel, H., 2009. Assessing neuronal networks: understanding AD. Prog. Neurobiol. 89 (2), 125–133.
Bonanni, L, Perfetti, B, Bifulchetti, S, Taylor, JP, Franciotti, R, Parnetti, L, Thomas, A, Onofri, M, 2015. Quantitative electroencephalographic utility in predicting conversion of MCI to DLB. Neurobiol. Aging 36 (1), 434–445.
Breiman, L, 2001. Random forests. Machine Learning 45 (1), 5–32.
Briels, CT, Stam, CJ, Scheltens, P, Bruins, S, Lues, I, Gouw, AA, 2020. In pursuit of a sensitive EEG functional connectivity outcome measure for clinical trials in AD. Clin. Neurophysiol. 131 (1), 88–95.
Brookes, MJ, Woolrich, MW, Barnes, GR, 2012. Measuring functional connectivity in MEG: a multivariate approach insensitive to source level leakage. Neuroimage 59 (4), 1910–1920.

Brun, A, Eckhorn, R, Jokert, H, Ebner, A, 2000. Amplitude envelope correlation detects coupling among irrelevant brain signals. Neuroreport 11 (7), 1509–1514.

Brun, M, Frederiksen, KS, Rhodus-Meester, HB, Barno, M, Gjerum, L, Koikkalainen, J, Uhremaa, T, Tolenon, A, van Gils, M, Tong, T, Guerrero, R, Rueschoot, D, Dyreomose, N, Andersen, BB, Simonens, AH, Lemstra, A, Halikainen, K, Mur, S, Herukka, SK, Remes, AM, Waldemar, G, Soininen, H, Mecocci, P, van der Flier, WM, Lötjönen, J, Hasselbalch, SG, 2019. Impact of a clinical decision support tool on dementia diagnostics in memory clinics: The PredictND validation study. Curr Alzheimer Res 16 (2), 91–101.

Brun, M, Rhodus-Meester, HF, Koikkalainen, J, Barno, M, Gjerum, L, Lemstra, AW, Barkhof, F, Remes, AM, Uhremaa, T, Tolenon, A, Rueckert, D, van Gils, M, Frederiksen, KS, Waldemar, G, Soininen, H, van der Flier, WM, 2018. Evaluating combinations of diagnostic tests to discriminate different dementia types. Alzheimers Dement 10, 509–518.

Caruana, R, Niculescu-Mizil, A, 2006. An empirical comparison of supervised learning algorithms. In: Proceedings of the Twenty-Third International Conference on Machine Learning (ICML ’06). Pittsburgh, Association for Computing Machinery, pp. 161–168.

Chen, W, Sahiner, B, Samuelsson, F, Pezeshk, A, Petrick, N, 2018. Calibration of medical diagnostic classifier scores to the probability of disease. Stat. Methods Med. Res. 27 (5), 1394–1409.

Cheny, D, Bostan, AC, Gaertz, W, Pang, EW, 2007. Event-related beamforming: a robust method for presurgical functional mapping using MEG. Clin. Neurophysiol. 118 (8), 1601–1704.

Colclough, GL, Woolrich, MW, Tewarie, PK, Brooks, MJ, Quinn, AJ, Smith, SM, 2016. How reliable are MEG resting-state connectivity metrics? Neuroimage 133, 63–75.

Dauw, M, van der Zande, JJ, van Dellen, E, Sommer, IE, Schel- tens, P, Lemstra, AW, Stam, CJ, 2016. Random forest to differentiate DLD from AD. Alzheimers Dement 12, 89–106.

De Haan, W, Pijnenburg, YA, Strijers, RL, van der Made, Y, van der Flier, WM, Schel- tens, P, Stam, CJ, 2009. Functional neural network analysis in frontotemporal dementia and AD using EEG and graph theory. BMC Neurosci. 10 (1), 1–12.

Dormann, CF, 2020. Calibration of probability predictions from machine-learning and statistical models. Global Ecol. Biogeogr. 29 (4), 760–775.

Douw, L, Niebro, D, Stam, CJ, Tewarie, P, Hillebrand, A, 2018. Consistency of magnetoencephalographic functional connectivity and network reconstruction using a template versus native MRI for co-localization. Hum Brain Mapp. 39 (1), 104–115.

Drummond, C, Holte, RC, 2003. C4.5: class imbalance, and cost sensitivity: why under-sampling beats over-sampling. In: Workshop on Learning from Imbalanced Datasets II, 1, Citeeseer, Washington DC, pp. 1–8.

Ehl, FM, Miller, BL, 2017. A clinicopathological approach to the diagnosis of dementia. Nat. Rev. Neurosci. 13 (8), 457.

Engels, M, Hillebrand, A, Van der Flier, WM, Stam, CJ, Schelten, P, Van Straaten, EC, 2016. Slowing of hippocampal activity correlates with cognitive decline in early onset AD. An MEG study with virtual electrodes. Front. Human Neurosci. 10, 238.

Engels, MMA, Van Der Flier, WM, Stam, CJ, Hillebrand, A, Schelten, P, Van Straaten, EC, 2017. Functional magnetic resonance imaging of the resting-state magnetoencephalography. Clin. Neurophysiol. 128 (8), 1426–1437.

Folstein, MF, Folstein, SE, McHugh, PR, 1975. Mini-mental state: a practical method for grading the cognitive state of patients for the clinician. J. Psychiat. Res. 12 (3), 189–198.

Francotti, R, Iacono, D, Della Penna, S, Pizzella, V, Torquati, K, Onofri, M, Romani, GL, 2006. Cortical rhythms reactivity in AD, LBD and normal subjects: a quantitative MEG study. 2015. Neurobiol. Aging 27 (8), 1100–1109.

Franchini, M, Demuara, M, Crobe, A, Marrosu, F, Stam, CJ, Hillebrand, A, 2016. The effect of epoch length on estimated EEG functional connectivity and brain network organisation. J. Neural Eng. 13 (3), 036015.

Garcés, P, Vicente, R, Wilbrath, M, Ángel Pinedo-Pardo, J, Eugenia López, M, Au- rentzen, S, Arcos, E, Marín de André, M, Yux, M, Sancho, M, Maestro, F, Fernández, A, 2013. Brain-wide slowing of spontaneous alpha rhythms in MCI. Front. Aging Neurosci 5, 100.

Gómez, C, Stam, CJ, Hornroth, R, Fernández, A, Maestú, F, 2009. Disturbed beta band functional connectivity in patients with MCI: an MEG study. IEEE Trans. Biomed. Eng. 56 (6), 1683–1690.

Gong, G, He, Y, Concha, L, 2009. Mapping anatomical connectivity patterns of human cerebral neural cortex using in vivo diffusion tensor imaging tractography. Cereb. Cortex 19 (3), 524–538.

Gouw, AA, Alvsema, AM, Tijms, BM, Borta, A, Schelten, P, Stam, CJ, van der Flier, WM, 2017. EEG spectral analysis as a putative early prognostic biomarker in nondemented, amyloid positive subjects. Neurobiol. Aging 57, 133–142.

Gouw, AA, Hillebrand, A, Schoonhoven, DN, Demuiri, M, Bis, P, Schelten, P, Stam, CJ, 2021. Routine magnetoencephalography in memory clinic patients: a machine learning approach. Alzheimers Dementia 13 (1), e12277.

Gouw, AA, Stam, CJ, 2016. Electroencephalography in the differential diagnosis of Alzheimer’s disease (pilot study). Hand, DJ, Till, RJ, 2001. A simple generalisation of the area under the ROC curve for multiple classification problems. Machine Learning 45 (2), 171–186.

Hillebrand, A, Barnes, GR, Bosboom, JL, Berendse, HW, Stam, CJ, 2012. Frequent-
1998. Frontotemporal lobar degeneration: a consensus on clinical diagnostic criteria. Neurology 51 (6), 1546–1554.
Ojala, M, Garriga, GC, 2010. Permutation tests for studying classifier performance. J. Mach. Learn. Res 11 (62), 1833–1863.
Onyike, CU, 2016. Psychiatric aspects of dementia. CONTINUUM 22 (2), 600.
Pedregosa, F, Varoquaux, G, Gramfort, A, Michel, V, Thirion, B, Grisel, O, Blondel, M, Prettenhofer, P, Weiss, R, Dubourg, V, Vanderplas, J, Passot, A, Cournapeau, D, Brucher, M, Perrot, M, Duchesnay, E, 2011. Scikit-learn: machine learning in Python. J. Mach. Learn. Res 12, 2825–2830.
Pijnenburg, YA, Strijers, RL, Van der Meade, Y, Van der Flier, WM, Scheltens, P, Stam, CJ, 2008. Investigation of resting-state EEG functional connectivity in frontotemporal lobar degeneration. Clin. Neurophysiol. 119 (8), 1732–1738.
Poil, SS, De Haan, W, Van der Flier, WM, Mansvelder, HD, Scheltens, P, Linkenaer-Hansen, K, 2013. Integrative EEG biomarkers predict progression to AD at the MCI stage. Front. Aging Neurosci. 5, 58.
Rascovsky, K, Hodges, JR, Knopman, D, Mendez, MF, Kramer, JH, Neuhaus, J, van Swieten, JC, Seaal, H, Dopper, EGP, Onyike, CU, Hillis, AE, Josephs, KA, Roeve, BF, Kertesz, A, Seeley, WW, Raskin, KP, Johnson, JK, Gorno-Tempini, M, Rosén, H, Priloueu-Latham, CE, Lee, A, Kipps, CM, Lillo, P, Piguet, O, Rohrer, JD, Rossor, MN, Warren, JD, Fox, NC, Galasko, D, Salmon, DP, Black, SE, Mesulam, M, Weintraub, S, Dickerson, BC, Diedhi-Schmid, J, Pasquier, F, Delacourte, V, Lebert, F, Pijnenburg, Y, Chow, TF, Manes, F, Graffman, J, Cappa, SF, Freedman, M, Grossman, M, Miller, BL, 2011. Sensitivity of revised diagnostic criteria for the behavioural variant of frontotemporal dementia. Brain 134 (9), 2456–2477.
Robison, SE, Mandell, AJ, 2014. Mutual information in a MEG complexity measure suggests regional hyper-connectivity in schizophrenic probands. Neuropsychopharmacology 40 (1), 251–252.
Sarica, A, Ceresa, A, Quattrone, A, 2017. Random Forest algorithm for the classification of neuroimaging data in AD: a systematic review. Front Aging Neurosci 9, 329.
Seiffert, C, Khoshgoftaar, TM, Van Hulse, J, Napoliolito, A, 2009. RUSBoost: A hybrid approach to alleviating class imbalance. IEEE Trans. Man Cybern. Syst. 40 (1), 185–197.
Sekihara, K, Nagarajan, SS, Poeppel, D, Marantz, A, 2004. Asymptotic SNR of scalar and vector minimum-variance beamformers for neuromagnetic source reconstruction. IEEE Trans. Biomed. Eng. 51 (10), 1726–1734.
Sheliha-Bahai, N, Sajadi, SA, Pierce, AL, 2017. Current role for biomarkers in clinical diagnosis of Alzheimer disease and frontotemporal dementia. Curr. Treat. Options Neurol. 19 (12), 46.
Slot, RE, Sikkes, SA, Berkhof, J, Brodaty, H, Buckley, R, Cavedo, E, Dardiotis, E, Guil-lo-Benarous, F, Hampel, H, Kochan, NA, Lista, S, Luck, T, Maruff, P, Malinu, JL, Kornhuber, J, Reisberg, B, Riedel-Heller, SG, Risacher, SL, Roehr, S, Sachdev, PS, Scarpace, N, Scheltens, P, Shluming, MB, Saykin, AJ, Verfaillie, SC, Visser, PJ, Vos, SJ, Wagner, M, Wolfgubner, S, Jessen, F, van der Flier, WM, 2019. SCD and rates of incident AD and non-AD dementia. Alzheimers Dement 15 (3), 465–476.
Steyerberg, EW, Vickers, AJ, Cook, NR, Gerds, T, Gonen, M, Obuchowski, N, Pencina, MJ, Kattan, MW, 2010. Assessing the performance of prediction models: a framework for some traditional and novel measures. Epidemiology 21 (1), 128.