Contents lists available at ScienceDirect

NeuroImage

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

A reusable benchmark of brain-age prediction from M/EEG resting-state signals

Denis A. Engemann[a,b,c,*], Apolline Mellot[b], Richard Höchenberger[b], Hubert Banville[b,d], David Sabbagh[b,e], Lukas Gemein[f,g], Tonio Ball[f,8], Alexandre Gramfort[b]

[a] Roche Pharma Research and Early Development, Neuroscience and Rare Diseases, Roche Innovation Center Basel, F. Hoffmann–La Roche Ltd., Basel, Switzerland
[b] Université Paris-Saclay, Inria, CEA, Palaiseau, France
[c] Max Planck Institute for Human Cognitive and Brain Sciences, Department of Neurology, D-04103, Leipzig, Germany
[d] Inserm, UMRS-942, Paris Diderot University, Paris, France
[e] Neuromedical AI Lab, Department of Neurosurgery, Medical Center – University of Freiburg, Faculty of Medicine, University of Freiburg, Engelbergerstr. 21, 79106, Freiburg, Germany
[f] Neurorobotics Lab, Computer Science Department – University of Freiburg, Faculty of Engineering, University of Freiburg, Georges-Köhler-Allee 80, 79110, Freiburg, Germany
[g] BrainLinks-BrainTools Cluster of Excellence, University of Freiburg, Freiburg, Germany
[h] InteraXon Inc., Toronto, Canada

ARTICLE INFO

Keywords:
Clinical neuroscience
Brain age
Electroencephalography
Magnetoencephalography
Machine learning
Population modeling
Riemannian geometry
Random forests
Deep learning

ABSTRACT

Population-level modeling can define quantitative measures of individual aging by applying machine learning to large volumes of brain images. These measures of brain age, obtained from the general population, helped characterize disease severity in neurological populations, improving estimates of diagnosis or prognosis. Magnetoencephalography (MEG) and Electroencephalography (EEG) have the potential to further generalize this approach towards prevention and public health by enabling assessments of brain health at large scales in socioeconomically diverse environments. However, more research is needed to define methods that can handle the complexity and diversity of M/EEG signals across diverse real-world contexts. To catalyse this effort, here we propose reusable benchmarks of competing machine learning approaches for brain age modeling. We benchmarked popular classical machine learning pipelines and deep learning architectures previously used for pathology decoding or brain age estimation in 4 international M/EEG cohorts from diverse countries and cultural contexts, including recordings from more than 2500 participants. Our benchmarks were built on top of the M/EEG adaptations of the BIDS standard, providing tools that can be applied with minimal modification on any M/EEG dataset provided in the BIDS format. Our results suggest that, regardless of whether classical machine learning or deep learning was used, the highest performance was reached by pipelines and architectures involving spatially aware representations of the M/EEG signals, leading to R^2 scores between 0.60-0.74. Hand-crafted features paired with random forest regression provided robust benchmarks even in situations in which other approaches failed. Taken together, this set of benchmarks, accompanied by open-source software and high-level Python scripts, can serve as a starting point and quantitative reference for future efforts at developing M/EEG-based measures of brain aging. The generality of the approach renders this benchmark reusable for other related objectives such as modeling specific cognitive variables or clinical endpoints.

Introduction

Aging-related disorders of the central nervous system affect hundreds of millions of patients, their caregivers and national health services. Over the past decades, important progress has been made in clinical neuroscience, resulting in improvements to clinical diagnosis and treatment (Walhovd et al. 2010; Ewers et al. 2011). Backed by increasingly advanced analytical methods, this has enabled fine-grained characterization of neurodegenerative conditions (Gaubert et al. 2019; Schumacher et al. 2021; Güntekin et al. 2021). Yet, from a public-health perspective, rather than focusing on pathology, it is essential to detect risk factors early within the general population to provide actionable feedback for preventive medicine, e.g., by targeting life-style changes. Such predictions are still challenging. Could it be helpful to look at biological rather than chronological age to better estimate the risk of declining brain health?

[*] Correspondence author.
E-mail address: denis.engemann@roche.com (D.A. Engemann).

https://doi.org/10.1016/j.neuroimage.2022.119521.
Received 14 December 2021; Received in revised form 4 July 2022; Accepted 25 July 2022
Available online 26 July 2022.

Recently, brain age has emerged as a concept for estimating biological aging in the general population (Cole and Franke 2017; Liem et al. 2017; Dosenbach et al. 2010). Biological aging can be inferred from the genome via telomere length, mitochondrial function, epigenetics and other cellular features (Ferrucci et al. 2020; Mather et al. 2011). Yet, the age of a person is only a noisy measure of these cellular processes (people of the same chronological age can have different biological ages). At the same time, biological aging affects brain structure and function (King et al. 2014), inducing loss of brain volume (Driscoll et al. 2009; Scahill et al. 2003) and characteristic changes in neuronal activity (Cabeza et al. 2002; Damoiseaux et al. 2008; Babiloni et al. 2006). A proxy of biological aging can, thus, be obtained by mapping chronological age to brain data from large populations of subjects using machine learning (Liem et al. 2017; Dadi et al. 2021). The resulting models can be used to compute an expectation of a person's age given her brain data. This is achieved by quantitatively comparing that person's brain data to the distribution of brain data across different ages within the general population. This statistical expectation can tell how old (or young) a brain "looks" (Spiegelhalter 2016), hence, predicting the risk of neurological complications potentially more precisely than the chronological age.

This empirical measure of biological aging derived from the general population has proven a useful marker of neurodegeneration and cognitive decline in clinical populations (Cole et al. 2018; Raffel et al. 2017; Denissen et al. 2021; Gonneaud et al. 2021). In these cohorts, patients typically appear to have older brains than their chronological age would suggest. Importantly, similar trends emerge when evaluating brain age in the general population where elevated brain age, compared to chronological age, has been associated with lower cognitive capacity, well-being, and general health (Dadi et al. 2021; Cole 2020; Wrigglesworth et al. 2021). Yet, so far, this approach has mainly been based on anatomical brain scans and hemodynamic signals obtained from magnetic resonance imaging (MRI). This limits the broad utility of brain age for public health, as cerebral MRI scans are usually collected when there is an indication, which can be too late. Even when people from the general population are motivated to participate in brain research, this only concerns a small fraction of society: MRI devices and neuroscientific studies are not equally accessible in all regions of the world and do not attract all people equally from within society, potentially leading to selection bias (Fry et al. 2017).

New hope to generalize this approach has been sparked by advances in large-scale modeling of biomedical outcomes from non-invasive electrophysiological data including magnetoencephalography (MEG) and electroencephalography (EEG) (Gaubert et al. 2019; Engemann et al. 2018). This line of research in clinical neurology may help develop assessments of brain health in many additional contexts in which MRI cannot be applied. First MEG-based brain-age models have allowed to validate MEG-derived brain age against MRI-derived brain age. Results from several studies have shown that the MEG- and MRI-derived brain aging asoimages are statistically related (Engemann et al. 2020; David Sabbagh et al. 2020; Xifra-Porxas et al. 2021). This overlap can be explained by electromagnetic field spread, independently of neuronal activity: As brain structure changes due to aging, cortical activity, even if unchanged, will project differently onto the M/EEG sensor array, making age indirectly decodable (Sabbagh et al. 2020). Importantly, multiple articles have found that neuronal activity captured by MEG adds specific information not present in MRI-derived brain age (Engemann et al. 2020; Xifra-Porxas et al. 2021), leading to improved prediction performance and richer neurocognitive characterization (Engemann et al. 2020).

While MEG can provide an important discovery context, it is unlikely to be the right instrument for addressing the availability issues of MRI-based brain age as MEG scanners are even rarer than MRI scanners. In this context, EEG can make a true difference as EEG is economical and allows for flexible instrumentation for neural assessments in a wide range of clinical and real-world situations including at-home assessments. First evidence suggests that MEG-based strategies for brain-age modeling can be translated to EEG. In an earlier publication (Engemann et al. 2020) we found that among many alternative features of varying data-processing complexity, the spatial distribution of cortical power spectra in the beta (13-30Hz) and alpha (8-13Hz) frequency band explained most of the MEG's performance as brain-age regressor. This type of information can be well accessed without source localization from the sensor-space covariance using spatial filtering approaches or Riemannian geometry (Sabbagh et al. 2020; Sabbagh et al. 2019), which has led to successful translation of this MEG-derived strategy to clinical EEG with around 20 electrodes (Sabbagh et al. 2020). In clinical and real-world contexts in which EEG is frequently collected, fine-grained spatial information may not be present as only a few electrodes are used. This has favored alternative EEG-derived brain-age models focusing on a wealth of spectral and temporal features (Al Zoubi et al. 2018) which may perform better on sparse EEG-montages and has enabled sleep-based brain age measures (Sun et al. 2019; Ye et al. 2020).

These results provide a sense of the flexibility and future potential of EEG-based brain age as a widely applicable real-world measure of brain health. Yet, to fully develop this research program, more and richer evidence is desirable. At this point, comparisons between different machine learning strategies are difficult. Most models were not only developed and validated in one specific context, but their implementations and data-processing routines are dataset-specific. Moreover, general machine learning approaches successful at pathology decoding should be well-suited for brain age modeling too, yet they have never been tested for that purpose (Gemein et al. 2020; Banville et al. 2020; Engemann et al. 2018). This makes it hard to know whether any strategy is globally optimal and where specific strategies have their preferred niche. As a result, uncertainty is added to comparisons between MEG, EEG and MRI, slowing down efforts of validating M/EEG-based brain age. Finally, to mitigate the impact of selection bias concerning the subjects investigated, it will be crucial to analyze many, socially and culturally diverse M/EEG datasets and find representations that are invariant to confounding effects that can raise issues of fairness and racial bias if remaining unaddressed (Choy, Baker, and Stavropoulos 2021). To develop the next generation of M/EEG-derived brain age models, to facilitate processing of larger numbers of diverse M/EEG-data resources and to avoid fragmentation of research efforts, standardized software and reusable benchmarks are needed.

In this paper we wish to make a first step in that direction. We provide reusable brain-age-prediction benchmarks for different machine learning strategies validated on multiple M/EEG datasets from different countries. Our benchmarks come in the form of readily usable, yet, easily adaptable Python scripts for computing brain age models and comparing their results. These scripts should not be taken as fully developed software but as practical templates for kick-starting future studies on brain age and biomarker learning on new datasets beyond the ones covered in this study. To facilitate reproducibility and usability, the benchmarks are built on top of well documented open-source software (MNE, PyRiemann, braindecode, scikit-learn) and The benchmarks are built on top of highly standardized dataset-agnostic code enabled by the BIDS standard (Gorgolewski et al. 2016; Niso et al. 2018; Appelhoff et al. 2019). This makes the benchmarks easy to extend in the future for additional datasets. The paper is organized as follows. The method section motivates the choice of the different machine learning benchmarks. The general data processing approach and software developed for this contribution are presented in the context of the benchmark. The selection of datasets is motivated, and datasets are then described in detail and compared regarding key figures that could provoke differences between benchmarks. Dataset-specific processing steps and peculiarities are highlighted. Then a model validation strategy is developed. The results section presents benchmarks on prediction performance across machine learning models and datasets and different performance metrics. The discussion inspects differences between models, modalities, and datasets, identifying unique niches, safe bets as well as unresolved challenges. The work concludes with practical suggestions on additional benchmarks that can be readily explored and extended by the community in future studies using the proposed tools and resources. The scripts and library code for this benchmark are publicly available on GitHub[1] and latest benchmark results (updated in real time) next to the config files specifying the analysis can be inspected on a dedicated website[2].

Methods

Brain age benchmarks

Many different approaches exist for ML in neuroscience, and it can be hard to select among them. The following categorization may help orient practical reasoning and study design. What varies in the taxonomy of methods discussed below is how much M/EEG data are statistically summarized before being presented to the learning algorithm. In other words, ML methods vary with respect to the extent to which compression and summary of the M/EEG signals is performed by the learning algorithm vs. feature-defining procedures performed before and independently of the machine learning algorithm.

A-priori defined, a.k.a. handcrafted, features

The first category represents approaches in which features are inspired by theoretical and empirical results in neuroscience or neural engineering. Here, M/EEG is summarized in a rigid fashion by global aggregation across sensors, time, and frequencies or by visiting specific regions of interest (Gemein et al. 2020; Sitt et al. 2014; Engemann et al. 2018). A meaningful composition of features requires prior knowledge of the (clinical) neuroscience literature, especially when interpretation of the model is a priority. In practice, it is convenient to extract all or the most relevant features discussed in a given field, apply multiple spatial and temporal aggregation strategies, and then bet on the capacity of the learning algorithm to ignore irrelevant features (Sitt et al. 2014). This motivates the use of tree-based algorithms like random forests (Breiman 2001) that are easy to tune, can fit nonlinear functions (higher-order interaction effects), and are relatively robust to the presence of uninformative features. As local methods that can be seen as adaptive nearest neighbors (Hastie et al. 2005), the predictions of random forests and related methods are bounded by the minimum and maximum of the outcome in the training distribution. For clinical neuroscience applications, this has proven to yield robust off-the-shelf prediction models that are relatively unaffected by noise in the data and in the outcome (Engemann et al. 2018). This approach is also a natural choice when using sparse EEG-montages with few electrodes.

Here we implemented a strategy pursued in (Gemein et al. 2020) and (Banville et al. 2020), aiming at a broad set of different summary statistics of the time-series or the power spectrum. This approach has turned out useful for a pathology detection task in which the labeling of EEG as pathological can be due to different clinical reasons, hence, affecting many different EEG signatures in potentially diffuse ways. Features were computed using the MNE-features package (Schiratti, Le Douget, Van Quyen, et al. 2018). More specifically we used as features (each computed for individual channels and concatenated across channels, and then averaged across epochs): the standard-deviation, the kurtosis, the skewness, the different quantiles (10%, 25%, 75%, 90%), the peak-to-peak amplitude, the mean, the power ratios in dB among all frequency bands (0 to 2Hz, 2 to 4Hz, 4 to 8Hz, 8 to 13Hz, 13 to 18Hz, 18 to 24Hz, 24 to 30Hz and 30Hz to 49Hz), the spectral entropy (Inouye et al. 1991), the approximate and sample entropy (Richman and Moorman 2000), the temporal complexity (Roberts, Penny, and Rezek 1999), the Hurst exponent as used in (Devarajan et al. 2014), the Hjorth complexity and mobility as used in (Päivinen et al. 2005), the line length (Esteller, Echauz, et al. 2001), the energy of wavelet decomposition coefficients as proposed in (Teixeira et al. 2011), the Higuchi fractal dimension as used in (Esteller, Vachtsevanos, et al. 2001), the number of zero crossings and the SVD Fisher Information (per channel) (Roberts, Penny, and Rezek 1999).

Covariance-based filterbank approaches

This category represents approaches in which the spatial dimension of M/EEG is fully exposed to the model, whereas temporal or spectral aspects of the signal are to some extent summarized before modeling. As M/EEG signals reflect linear superposition of neuronal activity projected to the sensors through linear field/potential spread, it is natural to use linear (additive) models for adaptively summarizing the spatial dimension of M/EEG signals (King et al. 2018; Stokes, Wolff, and Spaak 2015; King and Dehaene 2014). This intuition is driving the success of linear decoders for evoked response analysis but faces additional challenges when applied to power spectra (Sabbagh et al. 2020). Computing power features on M/EEG sensor-space signals renders the regression task a non-linear problem for which linear models will provide sub-optimal results (Sabbagh et al. 2019). In practice, this can be overcome by extracting nonlinear features like spectral power after anatomy-based source localization, or in a data-driven fashion that does not require availability of individual MRI scans. Spatial filtering techniques provide unmixing of brain sources based on statistical criteria without using explicit anatomical information, which has led to supervised spatial filtering pipelines (de Cheveigné and Parra 2014; Dähne et al. 2014). Another related strategy consists in computing features that are invariant to field spread. This can be achieved by Riemannian geometry, an approach first applied to M/EEG in the context of brain computer interfaces but that has also proven effective for biomarker learning (Barachant et al. 2012; Yger, Berar, and Lotte 2017; Rodrigues, Jutten, and Congedo 2019). These approaches have in common to favor the covariance of M/EEG sensors as a practical representation of the signals. Manipulating the covariance allows one to suppress the effects of linear mixing while, at the same time, exposing the power spectrum and the spatial structure of neuronal activity in each frequency band (Sabbagh et al. 2020). To scan along the entire power spectrum, one computes covariances from several narrow-band signals covering low to high frequencies (Sabbagh et al. 2020). This provides spatially fine-grained information of frequency-specific neuronal activity, hence the term *filterbank*.

Here we implemented the filterbank models from (Sabbagh et al. 2020; Sabbagh et al. 2019) based on Riemannian geometry that were found to provide a practical alternative to MRI-based source localization, although falling slightly behind in terms of performance. This may be explained by the model violations arising from computing the Riemannian embedding across multiple participants. The Riemannian embedding assumes can must linear field spread but each recording comes from a different head and different sensor locations, which, on the a few hum is explicitly modeled when computing individual-specific source estimates. It is an open question whether template-based source localization can improve upon the Riemannian pipeline, observing that in the case of MEG such a procedure would be informed by the head position in the MEG dewar. Both average brain templates and Riemannian embeddings mitigate field spread in a global way with the difference that the average template uses some anatomical information and approximate sensor locations in the context of MEG, whereas Riemannian embeddings are purely a data-driven procedure with some whitening based on the average covariance (across subjects).

To evaluate the benefit of a template-based anatomy, we included a filterbank model using source localization based on the *fsaverage* subject from FreeSurfer (Fischl 2012). The forward model was computed with a 3-layer Boundary Element Method (BEM) model. Source spaces were equipped with a set of 4098 candidate dipole locations per hemisphere. Source points closer than 5mm from the inner skull surface were excluded. The noise covariance matrices used along with forward solutions to compute minimum-norm estimates inverse operators

[1] https://github.com/meeg-ml-benchmarks/meeg-brain-age-benchmark-paper

[2] http://meeg-ml-benchmarks.github.io/brain-age-benchmark-paper

were taken as data-independent diagonal matrices. Diagonal values defaulted to the M/EEG-specific expected scale of noise (obtained via the “make_ad_hoc_cov” function from MNE-Python). All computations were done with MNE (Gramfort et al. 2014, 2013). For computational efficiency, source power estimates were obtained by applying the inverse operators to the subjects’ covariance data (MNE-Python function “apply_inverse_cov”). Dimensionality reduction was carried out with a parcellation containing 448 ROIs (Khan et al. 2018). This procedure closely followed the one from (Engemann et al. 2020), with the difference that here an MRI template was used instead of subject-specific MRIs. Finally, the 448 ROI-wise source power estimates represented as diagonal matrices were the inputs of the log-diag pipeline from (Sabbagh et al. 2020; Sabbagh et al. 2019). Features were computed using the coffeine package[3].

Deep learning approaches

This category concerns modeling strategies in which the outcome is mapped directly from the raw signals without employing separate a priori feature-defining procedures. Instead, multiple layers of nonlinear but parametric transformations are estimated end-to-end to successively summarize and compress the input data. This process is controlled by supervision and enabled by a coherent single optimization objective. In many fields, emerging deep learning methods keep defining the state of the art in generalization performance, often outperforming humans. Deep learning models are however greedy for data, and it may take hundreds of thousands if not millions of training examples until these models show a decisive advantage over classical machine-learning pipelines. Applied to neuroscience, where the bulk of datasets is small to medium-sized, deep learning models may or may not outperform classical approaches (Poldrack, Huckins, and Varoquaux 2020; Schulz, Thomas Yeo, et al. 2020; Roy et al. 2019; He et al. 2020). The success of using a deep-learning model may, eventually, depend on the amount of energy and resources invested in its development (Gemein et al. 2020).

Apart from high performance on standard laboratory M/EEG datasets and decoding tasks, deep learning models are attractive for other reasons. First, when very specific hypotheses about data generators or noise generators are available (Kietzmann, McClure, and Kriegeskorte 2019). In this setting, the model architecture can be designed to implement this knowledge, e.g. to explicitly extract band dressing power features in a motor decoding task. Second, these models have a strategic advantage when the data generating mechanism is not known at all, hence, few hypotheses about classes of features are available (Schirrmeister et al. 2017). In this setting, models with a generic architecture can learn and identify relevant features themselves without requiring expert knowledge of the researcher. With neural architecture search and automated hyperparameter optimization, there is also intense research to even reduce the amount of expert knowledge needed to create the network architecture itself. This flexibility has led neuroscientists to discover the framework as a vector for hypothesis-driven research probing brain functions and neural computation (Yamins and DiCarlo 2016; Bao et al. 2020). At the same time, this flexibility is equally beneficial under complex environmental conditions that degrade the quality of M/EEG recordings (e.g. real-world recordings outside of controlled laboratory conditions), in which the classes of relevant features are not a priori known and deep learning models can exploit the structure of the data and noise sources to provide robust predictions. (Banville et al. 2021).

Based on prior work, here we benchmarked two battle-tested general architectures (Gemein et al. 2020) implemented using the Braindecode package[4] (Schirrmeister et al. 2017; Gramfort et al. 2013). Braindecode is an open-source library for end-to-end learning on EEG signals. It is closely intertwined with other libraries. One of them is Mother of all BCI Benchmarks (MOABB) (Jayaram and Barachant 2018), which allows for convenient EEG-data fetching, MNE (Gramfort et al. 2013, 2014), implements well established data structures, preprocessing functionality, and more. A second key dependency is Skorch (Tietz et al. 2017), which implements the commonly known scikit-learn (Pedregosa et al. 2011) API for neural network training (Buitinck et al. 2013). For these reasons, Braindecode is equally useful for EEG researchers who desire to apply deep learning as well as for deep learning researchers who desire to work with EEG data. Braindecode builds on PyTorch (Paszke et al. 2019) and comprises a zoo of decoding models that were already successfully applied to a wide variety of EEG decoding classification and regression tasks, such as motor (imagery) decoding (Schirrmeister et al. 2017; Kostas and Rudzicz 2020), pathology decoding (Gemein et al. 2020; van Leeuwen et al. 2019; Tibor Schirrmeister et al. 2017), error decoding (Völker et al. 2018), sleep staging (Chambon et al. 2018; Perslev et al. 2021), and relative positioning (Banville et al. 2020).

For this benchmark and the task of age regression we used two Convolutional Neural Networks (ConvNets, sometimes abbreviated CNNs) (LeCun et al. 1999) namely ShallowFBCSPNet (BD-Shallow) and Deep4Net (BD-Deep) (Schirrmeister et al. 2017). BD-Shallow was inspired by the famous filter bank common spatial pattern (FBCSP) (Ang et al. 2008) algorithm. Initially, it has two layers that represent a temporal convolution as well as a spatial filter. Together with a squaring and logarithmic non-linearity it was designed to extract bandpower features. Of note, in the present context this architecture is closely related to SPoC (Dähne et al 2014) and, in therefore, in principle, has the capacity to deliver consistent regression models as was formally proven in previous work (Sabbagh et al 2020).

In contrast, BD-Deep is a much more generic architecture. In total, it has four blocks of convolution-max-pooling and is therefore not restricted to any specific features. While BD-Deep has around 276k trainable parameters and has therefore more learning capacity, BD-Shallow has only about 36k parameters.

It is important to note, that we did neither adjust the model architectures (apart from those changes required by the regression task) nor run task-specific hyperparameter optimization. Both ConvNets were used as implemented in Braindecode with hyperparameters that were already successfully applied to pathology decoding from the TUH Abnormal EEG Corpus (Gemein et al. 2020; van Leeuwen et al. 2019; Tibor Schirrmeister et al. 2017). For more information on Braindecode or the ConvNets, please refer to the original publication (Schirrmeister et al. 2017). For decoding, we converted the MEG input data from Tesla to Femtotesla, the EEG input data from Volts to Microvolts. Additionally, each dataset was rescaled separately by dividing each of its recordings by its global channel standard deviation (i.e., the standard deviation computed across all recordings), such that each dataset has roughly zero mean and unit variance (see Section Datasets).

General data processing strategy using BIDS and the MNE-BIDS pipeline

Neuroimaging and behavioral data are stored in many different complex formats, potentially hampering efforts of building widely usable methods, hence, impeding reproducible research. Our goal was to provide brain-age prediction models that can be directly applicable to any new electrophysiological dataset. For this purpose, we used the Brain Imaging Data Structure (BIDS) (Gorgolewski et al. 2016) which allows us to organize neuroimaging data in a standardized way supporting interoperability between programming languages and software tools. We used the MNE-BIDS software (Appelhoff et al. 2019) for programmatically converting M/EEG datasets into the BIDS format (Pernet et al. 2019; Niso et al. 2018). This has allowed us to access all datasets included in this work in the same way, enabling data analysis for all these datasets with the same code. We will now summarize the general workflow (cf. *Fig. 1*).

[3] https://github.com/coffeine-labs/coffeine

[4] https://braindecode.org

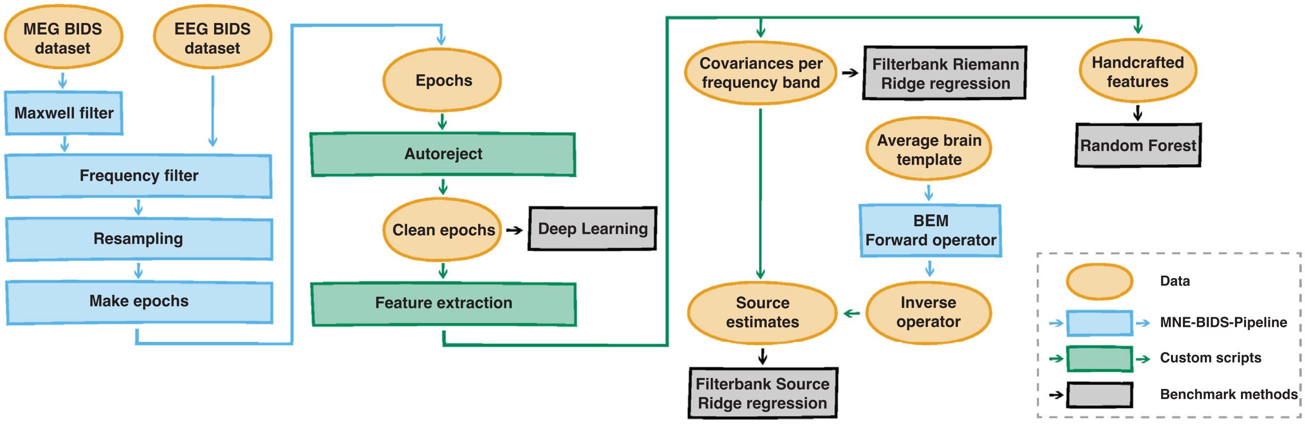

Fig. 1. Data processing, feature extraction and model construction based on the BIDS standard. This benchmark project provides a common data processing and feature extraction code allowing comparisons of different classical and deep learning-based machine learning models across different M/EEG datasets. Support for new datasets can be added with minimal modifications. For a detailed description consider the main text and the open-source code repository supporting this article (https://github.com/meeg-ml-benchmarks/meeg-brain-age-benchmark-paper).

For this study, we used the MNE-BIDS-Pipeline for automatic preprocessing of MEG and EEG data stored in BIDS format[5] (Jas et al. 2018). Its main advantage is that we can implement various custom analyses for different datasets without having to write any elaborate code. Modifying the overall processing pipeline or adapting a given pipeline to a new dataset only requires few edits. Controlling the pipeline is achieved through dataset-specific configuration files that specify the desired processing steps and options of the MNE-BIDS-Pipeline while dealing with the peculiarities of the data. The MNE-BIDS-Pipeline scripts themselves do not need to be modified and are readily applicable on diverse datasets.

We designed configuration files to implement data processing steps common to all datasets analyzed in this benchmark while handling dataset-specific details. Raw signals bandpass-filtered between 0.1 and 49Hz using a zero-phase finite impulse response (FIR) filter with Hamming window. Window length and transition bandwidth were automatically controlled by default settings of MNE-Python (v0.24). Like all preprocessing parameters, epoching can be easily adjusted to needs of a particular study via the dataset-specific config files. We considered epochs of 10-second length without overlap. These epochs coincided with eyes-closed or eyes-open resting-state conditions in some of the datasets. As additional channels measuring ocular and cardiac activity were not consistently available across datasets, we only implemented amplitude-based artifact rejection using the local autoreject method (Jas et al. 2017). Through 5-fold cross-validation, autoreject chose channel-specific rejection peak-to-peak-amplitude thresholds and then decided if a given epoch could be repaired using interpolation, or if it should be rejected to obtain clean data. We kept the default grid of candidate values for the hyperparameters 'rho' (the consensus proportion of bad channels leading to rejection of an epoch) and 'kappa' (maximum number of channels allowed to be interpolated). For 'rho' we considered a linearly spaced grid of 11 points between 0 and 1. For 'kappa' we considered 1, 4, or 32 channels. As the local autoreject is not yet supported in the MNE-BIDS pipeline, this step was implemented in a custom script (see the "compute_autoreject.py" in the code repository). This script could be easily edited to implement alternative artifact cleaning methods (e.g. RANSAC, Bigdely-Shamlo et al. 2015) or even omitted to probe the impact of preprocessing on model predictions. Apart from preprocessing, we also made use of the MNE-BIDS-Pipeline to generate forward solutions and inverse operators for the source localization approach based on template MRI (see section *Covariance-based filterbank approaches* for detailed explanations).

Each model of the benchmark is based on features extracted from clean epochs. Again, the conversion of datasets to BIDS has enabled feature extraction using one general script for all datasets ("compute_features.py" in the code repository).

Datasets

Large datasets and biobanks are the backbone of population modeling. In the past 10 years, this has led to a wealth of publications in cognitive neuroscience on modeling biomedical outcomes and individual differences in cognition from MRI data (Kernbach et al. 2018; Cole 2020; Smith et al. 2015). This has been enabled by consortia and large-scale institutional collaborations (Bycroft et al. 2018; Van Essen et al. 2013) that aim at recontextualizing existing data for open-ended future usage (Leonelli 2016). More recently, the first M/EEG datasets have emerged with a focus on characterizing populations (Taylor et al. 2017; Larson-Prior et al. 2013; Babayan et al. 2019; Obeid and Picone 2016; Niso et al. 2016; Valdes-Sosa et al. 2021; Bosch-Bayard et al. 2020). The selection of datasets for the present study did not aim at comprehensiveness but represents an attempt to secure a minimum degree of diversity. Social bias and fairness are important challenges, not only in the field of machine learning but also in biomedical research. It has been shown for modern biobanks that the sample deviates from the general population in important ways, oversampling Caucasian people with higher education degrees (Fry et al. 2017; Henrich and Heine 2010). For deployment of predictive biomarkers, this can have tragic consequences as clinical utility may depend on sex and ethnicity (Duncan et al. 2019). As a result, in EEG research, specific risks of racial bias have been recognized lately, highlighting the risk of selection bias and confounding, e.g., due to culture-specific hair style (Choy, Baker, and Stavropoulos 2021). Taken together, this emphasizes the importance of benchmarking on socially and culturally different datasets. Our selection includes M/EEG datasets from four different countries representing culturally and socioeconomically diverse contexts. To support construction of valid brain age models we focussed either on datasets sampled from the general population of healthy volunteers or on subsets of clinical EEG data that were labeled as non-pathological by medical experts. In the following we will provide a high-level introduction to the datasets, highlighting characteristic differences, challenges and opportunities for unique benchmarks.

Cam-CAN MEG data

The Cambridge Centre of Ageing and Neuroscience (Cam-CAN) dataset (Taylor et al. 2017; Shafto et al. 2014) has been the starting point of our efforts in building brain age models (Engemann et al. 2020;

[5] https://github.com/mne-tools/mne-bids-pipeline

Table 1
Aggregate cross-validation results across benchmarks and datasets.

Dataset	benchmark	$R^2_{(M)}$	$R^2_{(SD)}$	$MAE_{(M)}$	$MAE_{(SD)}$
Cam-CAN (MEG)	deep	0.63	0.11	8.74	1.23
Cam-CAN (MEG)	shallow	0.72	0.03	7.65	0.51
Cam-CAN (MEG)	filterbank-source	0.69	0.07	8.16	1.22
Cam-CAN (MEG)	filterbank-riemann	0.74	0.04	7.30	0.72
Cam-CAN (MEG)	handcrafted	0.49	0.06	10.68	1.00
Cam-CAN (MEG)	dummy	-0.02	0.03	15.90	1.22
LEMON (EEG)	deep	0.69	0.16	7.75	1.78
LEMON (EEG)	shallow	0.69	0.09	8.67	1.89
LEMON (EEG)	filterbank-source	0.67	0.11	8.67	1.07
LEMON (EEG)	filterbank-riemann	0.54	0.13	10.78	1.88
LEMON (EEG)	handcrafted	0.51	0.11	10.23	1.78
LEMON (EEG)	dummy	-0.13	0.17	18.70	1.60
CHBP (EEG)	deep	0.01	0.29	6.89	0.99
CHBP (EEG)	shallow	0.11	0.27	6.65	0.87
CHBP (EEG)	filterbank-source	-1.47	4.58	7.76	2.07
CHBP (EEG)	filterbank-riemann	-0.01	0.13	7.17	0.63
CHBP (EEG)	handcrafted	0.18	0.17	6.48	0.60
CHBP (EEG)	dummy	-0.04	0.05	7.33	0.83
TUAB (EEG)	deep	0.60	0.06	7.75	0.56
TUAB (EEG)	shallow	0.61	0.04	7.80	0.41
TUAB (EEG)	filterbank-source	0.56	0.06	8.43	0.57
TUAB (EEG)	filterbank-riemann	0.56	0.05	8.25	0.41
TUAB (EEG)	handcrafted	0.33	0.04	10.75	0.64
TUAB (EEG)	dummy	-0.01	0.01	13.55	0.82

Sabbagh et al. 2020) and we like to see it as a discovery context. The Cam-CAN dataset investigated healthy participants sampled from the general population without history of major disease (see exclusion criteria, Table 1 in Shafto et al. 2014). The combination of a wide, almost uniformly distributed age range and MEG data alongside MRI and fine-grained neurobehavioral results make it a rich resource for exploring aging-related cortical dynamics. On the other hand, models developed on this dataset may not be generalizable to real-world contexts in which EEG is operated. The following two sections are based on the methods description from our previous publications (Engemann et al. 2020; David Sabbagh et al. 2020).

Sample description. The present work was based on the latest BIDS release of the Cam-CAN dataset (downloaded February 2021). We included resting-state MEG recordings from 646 participants (female = 319, male = 327). The age of the participants ranged from 18.5 to 88.9 years with a mean age of 54.9 (female = 54.5, male = 55.4) and a standard deviation of 18.4 years. Data is provided in Tesla and has a standard deviation of 369.3 Femtotesla. We did not apply any data exclusion. Final numbers of samples reflect successful preprocessing and feature extraction. For technical details regarding the MEG instrumentation and data acquisition, please consider the reference publications by the Cam-CAN (Taylor et al. 2017; Shafto et al. 2014). In the following we highlight a few points essential for understanding our benchmarks on the Cam-CAN MEG data.

Data acquisition and processing. MEG was recorded with a 306 VectorView system (Elekta Neuromag, Helsinki). This system allowed measuring magnetic fields with 102 magnetometers and 204 orthogonal planar gradiometers inside a light magnetically shielded room. During acquisition, an online filter was applied between around 0.03Hz and 1000Hz. After bandpass filtering (0.1 - 49Hz), we applied decimation by a factor of 5, leading to a sample frequency of 200Hz (at the epoching stage). To mitigate the contamination of the MEG signal by environmental magnetic interference, we applied the temporal signal-space-separation (tSSS) method (Taulu, Simola, and Kajola 2005). Default settings were applied for the harmonic decomposition (8 components of the internal sources, 3 for the external sources) on a 10-s sliding window. To discard segments for which inner and outer signal components were poorly distinguishable, we applied a correlation threshold of 98%. As a result of this procedure, the signal was high pass filtered at 0.1Hz and the dimensionality of the data was reduced to 65, approximately. It is worthwhile to note that Maxwell filtering methods like tSSS merge the signal from magnetometers and gradiometers into one common low-rank representation. As a result, after tSSS, the signal displayed on magnetometers becomes a linear transformation of the signals displayed on the gradiometers. This leads to virtually identical results when conducting analyses exclusively on magnetometers versus gradiometers (Garcés et al. 2017). To reduce computation time, we analyzed the magnetometers for our benchmark. To deal with the reduced data rank, a PCA projection to the common rank of 65 was applied whenever the machine learning pipeline was sensitive to the rank (e.g., Riemannian filterbank models). For the full specification of the preprocessing, please refer to the “config_camcan_meg.py” file in the code repository.

LEMON EEG data

The Leipzig Mind-Body-Brain (LEMON) dataset offers rich multimodal EEG, MRI and fMRI data for a well characterized group of young and elderly adults sampled from the general population (Babayan et al. 2019). The LEMON dataset investigated healthy participants without history of major disease (see exclusion criteria, Table 1 in Babayan et al. 2019). As it was the case for the Cam-CAN data, here the research was conducted in a research context using high-end equipment accompanied by rich and fine-grained neurocognitive and behavioral assessments.

Sample description. EEG resting-state data from 227 healthy individuals from the LEMON dataset were included in this study. This sample contains 82 females (mean age = 44.2) and 145 males (mean age = 36), representing a clearly visible difference in the composition of the sample (*Fig.* 2). Their age distribution went from 20 to 77 years old with an average of 38.9 +- 20.3 years. Our sample covers the whole available dataset (downloaded September 2021) as we did not apply any exclusion criteria. It is a peculiarity of this dataset is that it is divided into 2 distinct age subpopulations, one between 20-35, the second between 55-77 (*Fig.* 2), rendering the mean a bad representation of the age distribution. Moreover, the public version of the datasets only provides ages in a granularity of 5 years to mitigate the risk of identifying participants. For the purpose of this study, we included the precise ages obtained through institutional collaboration. The impact on average modeling results turned out negligible, however. Data is provided in Volts and has a standard deviation of 9.1 Microvolts.

Data acquisition and processing. EEG was recorded with 62-channel active ActiCAP electrodes and a bandpass filter between 0.015Hz and 1kHz. We applied additional bandpass filtering between 0.1Hz and 49Hz. The channel placement implemented the 10-5 system (Oostenveld and Praamstra 2001). EEG data were sampled at 2500Hz. After bandpass filtering (0.1 - 49Hz), data were decimated by a factor of 5, yielding a final sampling frequency of 500Hz. As a peculiarity of the dataset, resting-state recordings encompass samples from two conditions: eyes closed and eyes open. Our pipeline explicitly respected these different conditions. To include a maximum of data and, potentially, a larger set of distinguishable EEG sources, we pooled the data prior to feature extraction. For the full specification of the preprocessing, please refer to the “config_lemon_eeg.py” file in the code repository.

CHBP EEG data

The Cuban Human Brain Mapping Project (CHBP) provides rich multimodal EEG and MRI data sampled from young to middle-aged adults from the general population (Valdes-Sosa et al. 2021; Hernandez-Gonzalez et al. 2011; Bosch-Bayard et al. 2020). The CHBP dataset investigated healthy participants without history of major disease (see exclusion criteria, Table 3 in Valdes-Sosa et al 2021). As for the Cam-CAN and LEMON data, research was carried out using high-end electrophysiological equipment in a biomedical research context. However, the data was collected in a Latin American mid-income country (Valdes-Sosa et al. 2021), adding a much-needed opportunity for increasing the diversity in population-level neuroscience datasets. This

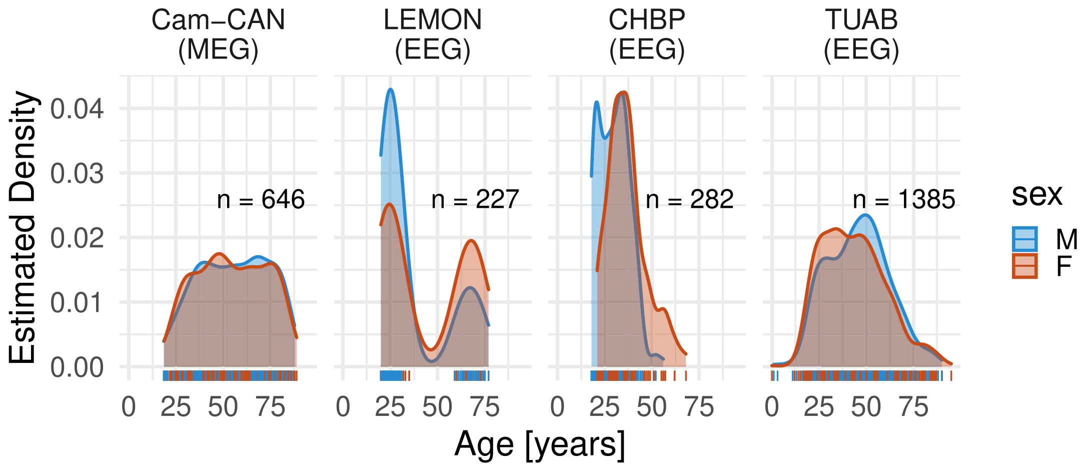

Fig. 2. Age distributions by gender by dataset. The kernel density (y axis) is plotted across the age range (x axis) for all four M/EEG datasets included in the study, separately for male (blue) and female (red) participants. Individual observations are displayed by rug plots at the bottom of each panel. The Cam-CAN data (MEG) show a wide age range with a quasi-uniform distribution and no obvious sex imbalance. This situation poses no a priori challenges for age prediction while, at the same time, analysis of MEG data may be more complex. The LEMON dataset included a group of young participants and a group of old participants, leading to a characteristic bimodal distribution. Sex imbalance is clearly visible with more male participants in the group of young participants and fewer male participants in the group of older participants. This may lead to potential sex differences in prediction success and renders the average age a bad summary of the age distribution. The CHBP data shows a rather reduced age range with a right-skewed age distribution and some sex imbalance (again more young male participants). Predicting the age can be expected to turn out more difficult on this dataset for the implied lack of density along the age range. Finally, the TUAB data present a symmetric age distribution with minor sex differences, however, a less uniform age distribution. This may lead to more pronounced errors in young and elderly participants. This may, however, be compensated for by the more generous sample size.

diversity expresses itself in the composition of EEG protocols which contain elements of real-world neurology exams, e.g., a hyperventilation task.

Sample description. EEG resting-state data from 282 healthy individuals from the CHBP dataset were included in this study. The sample contained 87 females (mean age = 36.7) and 195 males (mean age = 29.9), representing a clearly visible difference in the composition of the sample (*Fig.* 2). The overall age distribution went from 18 to 68 years with an average of 32 +/- 9.3 years. Data is provided in Volts and has a standard deviation of 6.6 Microvolts. Our sample covers the whole available dataset (download June 2021) as we did not apply any exclusion criteria. Final numbers reflect successful processing of the data.

Data acquisition and processing. EEG data were recorded using a MEDICID 5 system and two different electrode caps ofeither 128 or 64 channels with a subset of 53 common channels. The channel placement implemented the 10-5 system (Oostenveld and Praamstra 2001). Here we focused the analysis on the subset of common channels present in all recordings, leading to 53 channels. We applied additional bandpass filtering between 0.1Hz and 49Hz. As in the LEMON dataset, resting-state recordings encompassed samples from eyes-closed and eyes-open conditions. Again, we pooled both conditions prior to feature extraction. Note that for the data release (downloaded July 2021) used in this work, we could not benefit from the expert-based annotations of clean data. The results obtained on this dataset may therefore be impacted by quality issues to unknown extents.

For the full specification of the preprocessing, please refer to the "config_chbp_eeg.py" file in the code repository.

TUAB EEG data

The Temple University Hospital Abnormal EEG Corpus (TUAB) provides socially and ethnically heterogeneous clinical EEG data (Obeid and Picone 2016) mostly from Latin-American and African American participants (personal communication, Joseph Picone). As a peculiarity, the EEG data is obtained from an archival effort of recovering different EEG exams from the Temple University Hospital in Philadelphia. EEGs were administered for different purposes and indications and subsequently labeled as pathological or non-pathological. The clinical and social diversity render the TUAB dataset an important resource for electrophysiological population modeling (Gemein et al. 2020; Sabbagh et al. 2020).

Sample description. Here, we focused exclusively on the EEG recordings labeled as not pathological by medical experts comprising a subsample of 1385 subjects (female = 775 and males = 610). This sample contained individuals ranging from newborn children (min age = 0 for female and min age = 1 for male) to elderly (max age = 95 for female and 90 for male) people (*Fig.* 2). The average age is 44.4 +/- 16.5 years. Data is provided in Volts and has a standard deviation of 9.7 Microvolts. The data processing closely followed our previous work on the TUAB data (Sabbagh et al. 2020). For further details about the dataset, please refer to the reference publications (Harati et al. 2014; Obeid and Picone 2016).

Data acquisition and processing. EEG data were recorded using different Nicolet EEG devices (Natus Medical Inc.), equipped with between 24 and 36 channels. For channel placement, the 10-5 system was applied (Oostenveld and Praamstra 2001). EEG data were sampled at 500Hz. After bandpass filtering (0.1 - 49Hz), data were resampled to 200Hz. All sessions have been recorded with an common reference. Here we considered a subset of 21 common channels. As channel numbers differed across recordings, re-referencing was necessary. For consistency, we also applied re-referencing with an average reference on all other EEG datasets. As sampling frequencies were inconsistent across recordings, we resampled the data to 200Hz. For many patients, multiple recordings were available. For simplicity we only considered the first recording. For the full specification of the preprocessing, please refer to the "config_tuab_eeg.py" file in the code repository.

Model evaluation and comparison

To gauge model performance, we first defined a baseline model that should not provide any intelligent prediction. As in previous work (Sabbagh et al. 2020; Sabbagh et al. 2019; Engemann et al. 2020), we employed a dummy regressor model as a low-level baseline in which the outcome is guessed from the average of the outcome on the training data. This approach is fast and typically converges with more computationally demanding procedures based on permutation testing that we shall briefly outline below. Of note, our benchmark code can be used to run and display any alternative user-defined baseline, e.g., by passing alternative csv files for the dummy model.

This is particularly relevant for the present benchmark where the combinatorial matrix of machine learning models (including deep learning) versus datasets would lead to unpleasant computation times when applying tens of thousands of permutations. The same can be said for other approximations focusing on ranking statistics across hundreds of Monte Carlo cross-validation iterations (Sabbagh et al. 2019). Finally, another approach relies on large left-out datasets, entirely independent from model construction, in which predictions can be treated like random variables, hence, classical inferential statistics are valid. In previous work (Dadi et al. 2021), permutation tests and the non-parametric bootstrap were employed on more than 4000 left-out data points to assess performance above chance and pairwise differences between models. Such generous held-out datasets are not available in the present setting, nor can we readily compute statistics across folds, as cross-validation iterations are not statistically independent. We therefore implemented a less formal approach comparing competing models against dummy regressors and against each other based on standard 10-fold cross-validation based on fixed random seeds. This ensured that for any model under consideration, identical data splits were used. Of note, our reusable benchmark code allows interested readers to implement more exhaustive model comparison strategies.

For scoring prediction performance, we focused on two complementary metrics. The coefficient of determination (R^2) score – “bigger is better” interpretation – and the mean absolute error (MAE) – “smaller is better” interpretation. Considering the dummy regressor, the R^2 score is a natural choice as it quantifies the incremental success of a model over a regressor returning the average of the training-data as a guess for the outcome. Compared to Pearson correlations that are sometimes used in applied neuroscience studies, the R^2 metric is more rigorous as it is sensitive to the scale of the error and the location: Predictions that are entirely biased, e.g, shifted by a large offset, could still be correlated with the outcome. In contrast, the R^2 metric clearly penalizes systematically wrong predictions by assigning scores smaller than 0. Positive predictive success thus falls into a range of R^2 between 0 and 1 (higher scores are better). This facilitates comparisons across models within the same dataset while posing challenges when comparing models across datasets.

We therefore considered the MAE which has the benefit of expressing prediction errors at the scale of the outcome. This is particularly convenient for scientific interpretation when the outcome has some practical meaning as is the case in the present benchmarks on age prediction (smaller scores are better). Importantly, the MAE does not per se resolve the problem of comparisons across datasets as the meaning of errors entirely depend on the distribution of the outcome: Small errors in years are good for datasets with wide age distributions but bad in datasets with narrow age distributions. This obviously calls for contextualizing the MAE against a dummy baseline regression model. While this does not necessarily facilitate comparisons across datasets, it helps make visible situations in which one cannot rely solely on the R^2 for model comparisons.

For model comparisons, we employed a multivariate extension of Bland-Altman plots (Bland and Altman 1999; Möller et al. 2021). Following Möller et al., the lack of agreement threshold was computed from the the chi-square quantile function and the standard deviation across the models $\chi^2(0.95,\ m-1)\ \frac{1}{\sqrt{(m-1)}}\ n^{-1}\ \sum_i^n s_i$ where m is the number of models and s is the standard deviation across models.

Computational considerations and software

M/EEG data processing. BIDS conversion and subsequent data analysis steps were carried out in Python 3.7.1, the MNE-Python software (v0.24, Gramfort et al. 2014, 2013), the MNE-BIDS package (v0.9, Appelhoff et al. 2019) and the MNE-BIDS-pipeline on a 48-core Linux high-performance server with 504 GB RAM. The joblib library (v1.0.1) was used for parallel processing. For artifact removal, the latest development version (v0.3dev) of the autoreject package (Jas et al. 2017) was used.

Classical machine learning benchmarks. For future computation, the mne-features (0.2, Schiratti, Le Douget, Le Van Quyen, et al. 2018, PyRiemann (v0.2.6) and the coffeine (0.1, David Sabbagh et al. 2020) libraries were used. Analyses were composed in custom scripts and library functions based on the Scientific Python Stack with NumPy (v1.19.5, Harris et al. 2020), SciPy (v1.6.3, Virtanen et al. 2020) and pandas (v.1.2.4, McKinney and Others 2011). Machine-learning specific computation was composed using the scikit-learn package (Pedregosa et al. 2011). Analysis was carried out on a 48-core Linux high-performance server with 504 GB RAM. Model training and evaluation completed within a few minutes to hours. However, feature computation could last several days, depending on the dataset and the types of features.

Deep learning benchmarks. A high-performance Linux server with 72 cores, 376 GB RAM and 1 or 2 Nvidia Tesla V100 or P4 GPUs was used. Code was implemented using the PyTorch (Paszke et al. 2019) and braindecode (Schirrmeister et al. 2017) packages. Model training and evaluation completed within 2-3 days.

Data visualization. Graphical displays and tables were composed on an Apple Silicon M1 Macbook Pro (space gray) in R (v4.0.3 “Bunny-Wunnies Freak Out”) using the ggplot2 (v3.3.5, Wickham 2011), patchwork (v1.1.1, Pedersen 2019), ggthemes (v4.2.4) and scales (v1.1.1, Arnold 2017) packages with their respective dependencies.

Results

For the age prediction benchmark, we considered five alternative approaches: heterogeneous handcrafted features & random forest (‘handcrafted’), filterbank features based on Riemannian embeddings & ridge regression (‘filterbank-riemann’), filterbank features based on source localization with MRI-average template & ridge regression (‘filterbank-source’), a shallow deep learning architecture (“shallow”) and a 4-layer deep-learning architecture (‘deep’). These approaches were benchmarked across four M/EEG datasets: The Cambridge Centre of Ageing and Neuroscience (Cam-CAN) dataset (Taylor et al. 2017), the Cuban Human Brain Mapping Project (CHBP) dataset (Valdes-Sosa et al. 2021), the Leipzig Mind-Brain-Body (LEMON) dataset (Babayan et al. 2019) and the Temple University Hospital Abnormal EEG Corpus (TUAB) dataset (Obeid and Picone 2016). Generalization performance was estimated using 10-fold cross validation after shuffling the samples (fixed random seed). The coefficient of determination (R^2, bigger is better) was used as a metric enabling comparisons between datasets independently of the age distribution, mathematically quantifying the additional variance explained by predicting better than the average age. A dummy model empirically quantifies chance-level prediction by returning the average age of the training data as prediction. The results are displayed in Fig. 3. One can see that on most of the datasets all machine learning models achieved R^2 scores well beyond the dummy baseline. The highest scores were observed on the Cam-CAN MEG dataset, followed by the LEMON EEG dataset. Caution is warranted though to avoid premature conclusions: The R^2 offers a common scale that explicitly compares the incremental model performance over the average predictor. This is achieved by dividing the sum of squares of the model’s prediction by the sum of squares of the average predictor but, in turn, depends on the distribution of age. As a result, this can be misleading in cross-dataset comparisons when the variance of the outcome is not the same, which is the case here (cf. Fig. 2). We therefore also computed results using the mean absolute error (MAE, smaller is better) as a performance metric (Fig 4). One can now see that the overall distribution of scores, including the scores of the dummy model, depend not only on the dataset but also on its age range. Where the range is small, improvements over the baseline models are harder to observe. Moreover, comparing MAE scores across datasets without taking into account the baseline can yield misleading

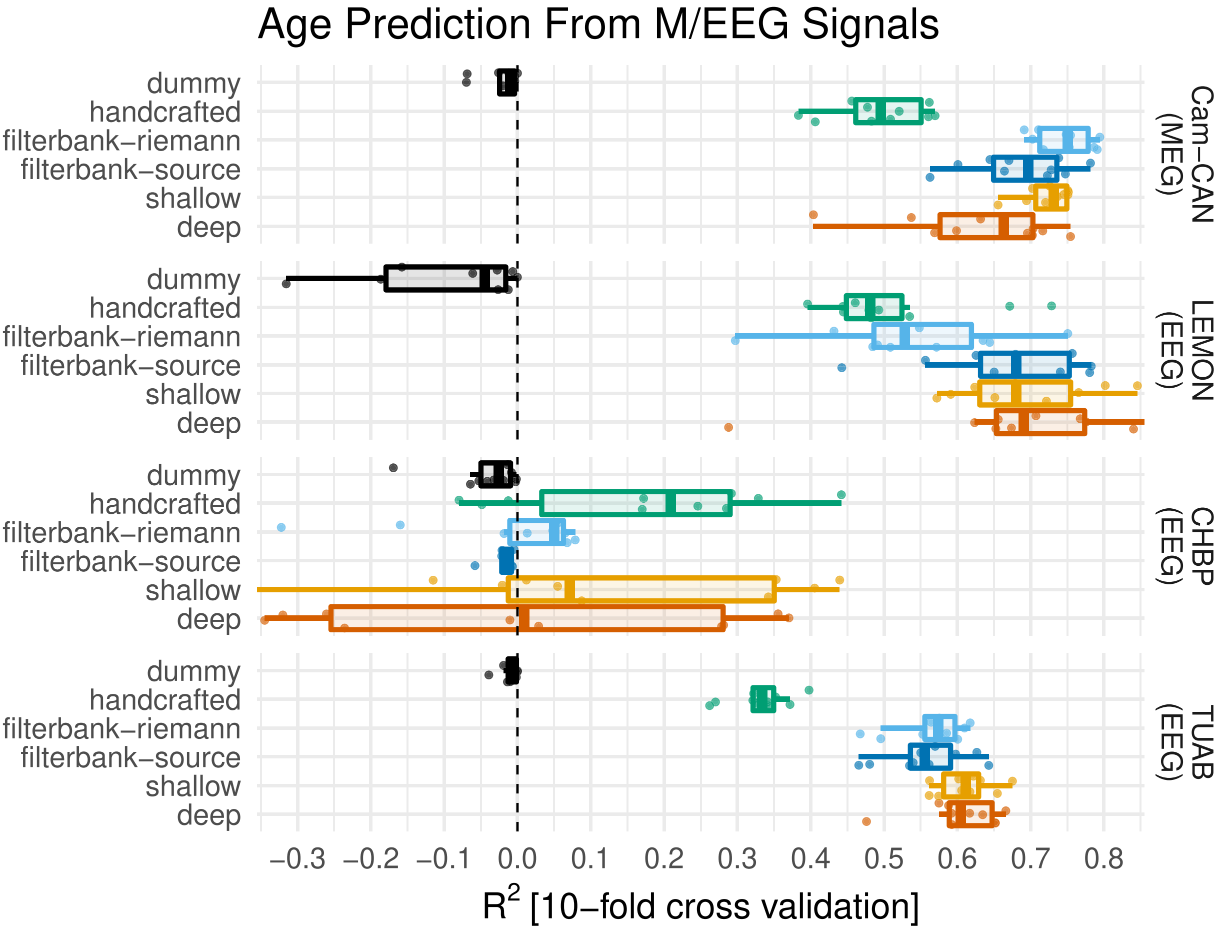

Fig. 3. Age prediction benchmarks across M/EEG datasets (R^2 score). Generalization performance was assessed by 10-fold cross-validation and the R^2 score (bigger is better) for five machine learning strategies compared against a dummy model (rows) and four datasets (panels). Across datasets, dummy models were mostly well-calibrated with R^2 scores close to zero. The LEMON dataset was one exception as dummy scores were systematically worse than chance, which can be explained by the bimodal age distribution (cf. Fig. 2), rendering the average age a bad guess for the age. The 'handcrafted' benchmark delivered moderate but systematic prediction success across all datasets. The two filterbank models performed well across datasets with similar performances, markedly higher than for the 'handcrafted' approach. The only exception was the CHBP benchmark for which neither the filterbank nor the deep models delivered useful predictions. Note that here, for the 'filterbank-source', a single fold with an abysmal R^2 score of -15 was obtained (x limits constrained to a range between -.3 and 1.0). Overall, the deep learning benchmarks performed similarly to the filterbank models.

conclusions. For example, the same score of *e.g.* an MAE = 10 can be way above chance in one dataset (Cam-CAN) but below chance in another dataset (CHBP). To alleviate this problem, normalized MAE scores have been suggested in which the MAE scores are related to the range of the age distribution (James H. Cole, Franke, and Cherbuin 2019). This does not come without its own problems, as then outliers in non-uniform distributions could drive the scores. As research keeps evolving on this topic and the community has not yet agreed on the best metric, we recommend considering multiple classical machine learning metrics when comparing model performance – in critical awareness of their respective limitations.

Confronting the relative performances of models to the dummy baseline in Fig. 3 and Fig. 4, one can see overall similar performance rankings between the models, regardless of the metric. See Table 1 for side-by-side comparisons of the aggregated cross-validation distributions. The big-picture results argue in favor of the importance of fine-grained spatial features for M/EEG prediction while considering important between-dataset heterogeneity. Both filterbank pipelines provide features based on spatially aware representations of the M/EEG signals, which either explicitly or implicitly deal with the spatial spread of electrical potentials and fields characteristic for M/EEG signals. The source-level filterbank approximates source localization using the average MRI template, whereas Riemannian embeddings provide non-linear spectral features that are affine invariant, hence, independent of linear mixing. The deep benchmarks, on the other hand, implied spatial-filtering layers capable of mimicking source localization by learning an unmixing function. Surprisingly, using the average MRI template instead of the Riemannian embedding to construct a filterbank model did not lead to consistent improvements across datasets, suggesting that both approaches may be equally effective in practice. We would have conjectured that even an imprecise biophysical head model would provide inverse solutions leading to more accurate unmixing of M/EEG sources. Compared to our previous benchmarks (Engemann et al. 2020; David Sabbagh et al. 2020) favoring filterbank models based on source-localization, one has to point out that this finding may reflect at least two differences: The use of an MRI template instead of individual co-registration and the use of empty-room-based suppression of environmental noise. The second factor may be less relevant for EEG though where empty room recordings are not available and data-based covariances are more common in event-related studies where brain activity induced by stimuli is compared against the background resting-state activity. As a practical implication, and if inspection of the brain sources

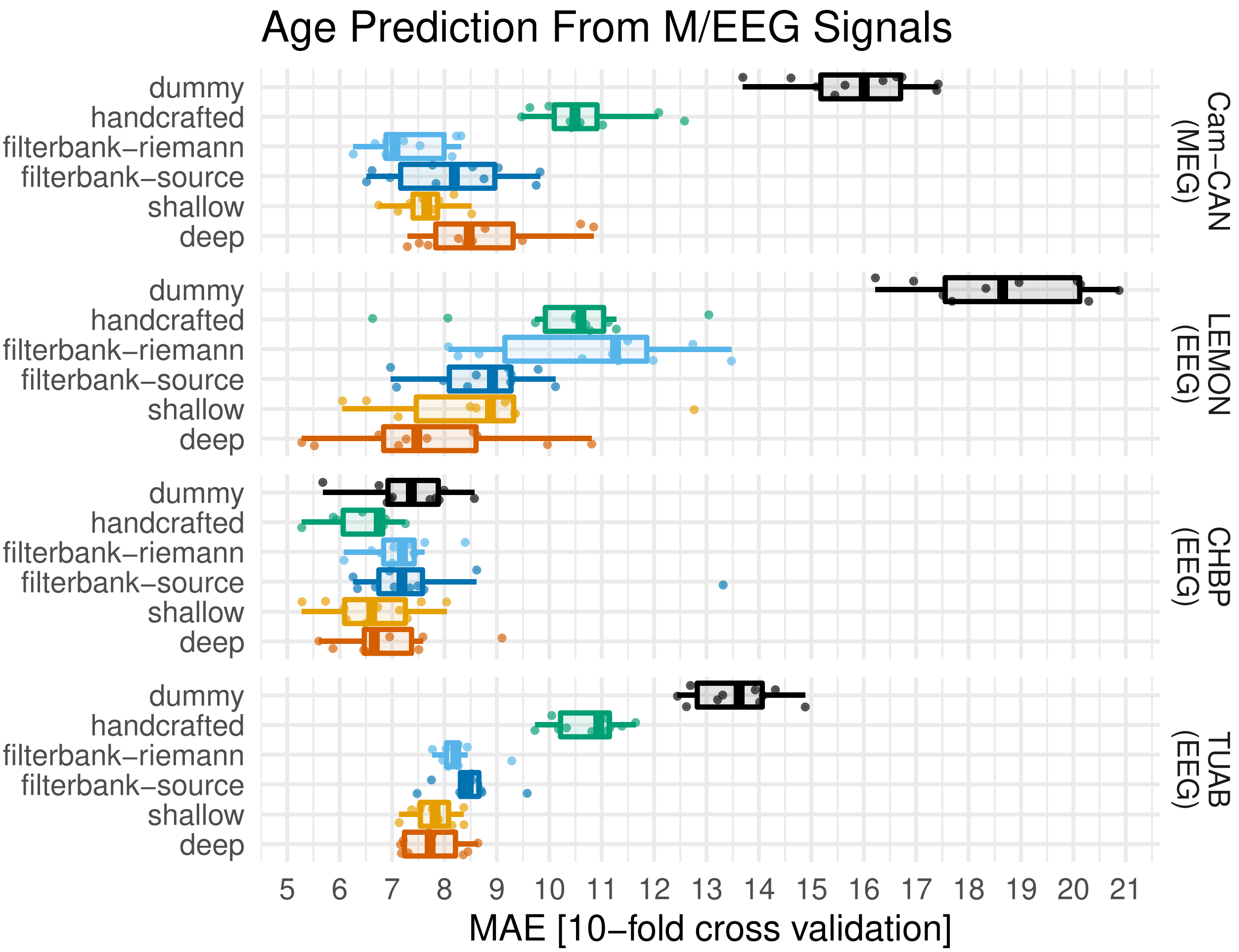

Fig. 4. Age prediction benchmarks across M/EEG datasets (mean absolute error). Same visual conventions as in Fig. 3. As the MAE (smaller is better) is sensitive to the scale and distribution of the outcome, one can see characteristic differences across datasets. The distribution of the dummy scores provides an estimate of the random guessing. As before, in all but the Cuban datasets all benchmarks achieved MAE scores markedly better than the dummy with no overlap between model and dummy distributions. Model rankings resemble the ones obtained using the R^2. On the LEMON data, the deep benchmark now presented a slight advantage over all other benchmarks.

is not a priority, the purely data-driven pipelines may be more practical as no additional MRI-based data processing is needed (cf. Fig. 1).

Interestingly, none of these approaches involving spatially fine-grained representations of the M/EEG signal worked well on the CHBP data, whereas the random forest on top of hand-crafted features scored systematically better than the dummy baseline. This may be related to three factors that come together in the CHBP benchmark dataset: Like the LEMON dataset, the sample size is relatively small. Second, the age distribution is far less uniform, leading to underrepresentation of elderly participants. This makes the learning task at hand harder as models have fewer training examples from elderly populations. These challenges apply equally to all machine learning benchmarks, hence, do not explain why the random forest model on hand-crafted features is working to some extent. In this context, it may be worthwhile to remember that the CHBP uses two different EEG montages, one with 128, one with 64 electrodes (both implementing 10-05 electrode placement). Despite only focussing on the 53 common channel locations, the larger VS higher number of electrodes may induce substantial differences in the covariance structure of the signals (Nunez and Srinivasan 2006) as the same electrodes in the context of a smaller electrode array implies more spatial averaging of cortical activity. This may have affected the random-forest pipeline less strongly as the hand-crafted features extracted marginal channel-wise summary statistics of the time-series or the power spectrum rather than pairwise interactions. Progress on this specific benchmark may therefore involve explicit consideration of the montage when selecting samples for cross-validation or even at the level of the machine learning model (e.g., by including the number of electrodes or montage type as covariate). Moreover, future availability of samples from older populations in the CHBP dataset will help disambiguate this point. Finally, once the expert-based quality-control annotations are considered for epochs-selection, the results obtained in this benchmark may change (see section Datasets/CHBP EEG data for details).

A different type of challenge is illustrated by the benchmarks on the LEMON dataset. As the age distribution is bimodal here (Fig. 2), the R^2 score is not well calibrated as the average predictor will not provide a reasonable summary of the distribution. This is not automatically mitigated by considering the MAE as a metric. On the other hand, it will not affect the ranking of the machine learning models, which compare overall well to results obtained on the Cam-CAN and the TUAB datasets. To obtain a more rigorous baseline, one could envision a group-wise average predictor that, depending on the age group, would return the groups' respective average age from the training data. We did not implement such a custom baseline here as it was our goal to stick to standard routines provided by the software libraries our benchmarks were based on. Second, it was our intention to expose such issues as this may stimulate future research and development.

When applying brain age models for research or clinical purposes to characterize individuals, e.g., correlating brain age with clinical scores,

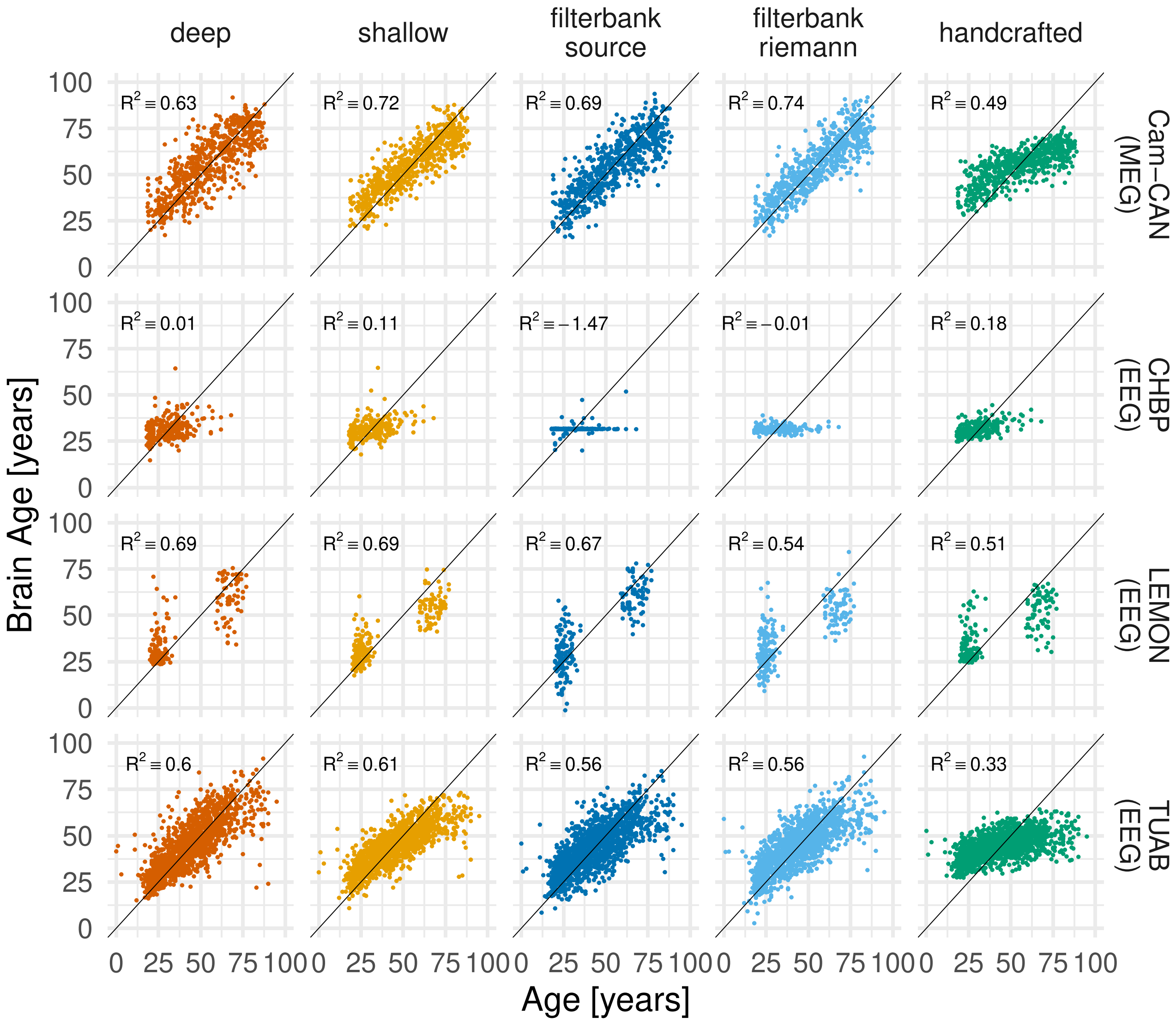

Fig. 5. Scatter plots of actual versus brain-predicted age. Cross-validated age predictions were generated by concatenating the predictions on the 10 held-out cross-validation splits. Every point represents one participant. For convenience, previously reported average R^2 scores are displayed in each panel. Diagonal identity lines represent ideal error-free predictions.

it is important to account for residual correlations with chronological age. This topic has been recently discussed as brain age bias expressed in systematic errors in younger or older age groups, which may arise from the training distribution, regularization or limited expressiveness of the model. Fig. 5 displays individual-level correlations between chronological age and brain-predicted age. One can readily see that better performing benchmarks distributed more narrowly around the identity line (e.g. shallow net in Cam-CAN), whereas worse performing models were systematically off (e.g. handcrafted in TUAB), inducing residual correlations between brain age and age. Also when focussing on the younger and older age ranges in the best-performing benchmarks, over prediction and under prediction can be spotted by the naked eye. The results emphasize the importance of deconfounding or residualizing for chronological age as a key practical issue when applying the presented benchmarks (see Smith et al. 2019; Liang, Zhang, and Niu 2019 for general discussion, Engemann et al 2020 for applied examples with MEG). As for applications of MRI-derived brain age, taking into consideration the residual error structure remains an important issue for applications of M/EEG-derived brain age. It will be worthwhile to investigate if bias-correction methods focusing on regression to the mean can improve the presented benchmarks (Liang, Zhang, and Niu 2019).

The overall similar associations between brain age and chronological age (Fig. 5) pose the question about the redundancy and complementary the model predictions. To better understand the relationship between the models, we analyzed their predictions on a sample-by-sample basis using a multivariate Bland-Altman plot (Bland and Altman 1999; Möller et al. 2021). Results in Fig. 6A show that, except for an important outlier in the CHBP dataset, model agreement was rather uniformly distributed along the age range of each dataset. Nevertheless, increased mismatch between the models can be spotted in younger and older age groups, potentially enabling improvements through model averaging. This raises the question if pooling the models using simple bagging (Breiman 1996) can improve model performance by capturing potentially informative differences between models. Results revealed that simple model averaging via bagging did not lead to clear improvements over the best model but, roughly, preserved the best model's performance (Fig. 6B). This may suggest that the models learned overall similar functions, potentially tuning into similar aspects of the input data.

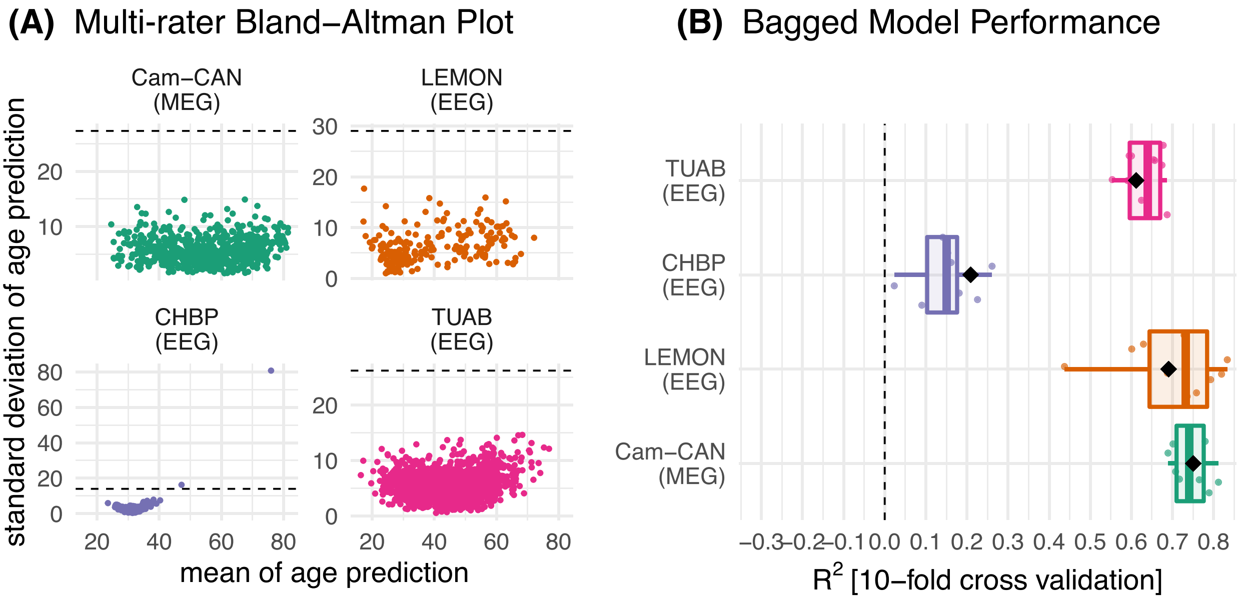

Fig. 6. Model redundancy and complementarity. Panel (A) plots the sample-wise average over the prediction from different benchmarks against the standard deviation over the benchmarks, implementing a multi-rater Bland-Altman plot. The dashed line represents the statistically expected lack-of-fit threshold. Panel (B) shows performance after averaging model predictions prior to scoring (R^2). For convenience, the median performance of the best model is indicated by black diamonds.

Discussion

In this study, we proposed empirical benchmarks for age prediction comparing distinct machine learning approaches across diverse M/EEG datasets comprising, in total, more than 2500 recordings. The benchmarks were implemented in Python based on the MNE-software ecosystem, the Braindecode package and the BIDS data standard. The explicit reliance on the BIDS standard renders these pipelines applicable to any M/EEG data presented in the BIDS format. This enabled coherent side-by-side comparisons of classical machine learning models and deep learning methods across M/EEG datasets recorded in different research or medical contexts.

Our cross-dataset and cross-model benchmarks pointed out stable ranking of model performance across two metrics, the R^2 score and mean absolute error (MAE). R^2 scores have been less consistently reported in the literature, however, the top MAE scores observed across datasets in this benchmark of 7 to 8 years are well in line with reports from previous publications (Sun et al. 2019; Sabbagh et al. 2020; Xifra-Porxas et al. 2021). While direct comparisons against MRI were not performed in this study, the present benchmarks would be compatible with the impression that for what concerns the overall performance of age prediction, M/EEG features are slightly weaker than MRI features (Engemann et al. 2020; Xifra-Porxas et al. 2021). We found that, overall, Riemannian filterbank models and deep learning models achieved the best scores (highest R2 and lowest MAE). On the other hand, random forests based on hand-crafted features delivered robust performance in the sense that performance was never at the top but still present when other methods failed. This can be explained by the fact that the hand-crafted features – compared to the other models – did not handle field-spread and volume-condition effects, potentially leading to lower performance with small datasets, while, at the same time, the random forest is robust in the sense of returning predictions from within the support of the training distribution, preventing extrapolation errors.

In line with previous work (Gemein et al. 2020), these results suggest that deep learning methods do not necessarily show a consistent advantage over classical pipeline models: Similar performance may be explained by the fact that our filterbank models and the deep models imply similar spatially aware representations of the M/EEG data (see results section for detailed discussion in context). Of note, the shallow and the deep model (both described in Schirrmeister et al 2017) were on par, suggesting that the additional convolutional layers of the deep net, here, did not add useful model complexity. Moreover, given the relatively small training datasets, it can be considered good news that these parameter-rich models did not seem to overfit as was evidenced by comparisons against simpler classical models. Yet, it may be simply a matter of collecting larger samples until deep learning approaches may reveal their advantage at extracting more elaborate representations of M/EEG signals. While concrete estimates for the decisive sample size for M/EEG are not available, related research on the scaling of machine learning models applied to MRI in the UK biobank would suggest that one would need sample sizes in the order of ten to a hundred of thousands if not one million (Schulz, Yeo, et al., 2020; Schulz et al. 2022). This may lead to positioning M/EEG-based brain age prediction on par with MRI-based brain age prediction just as MRI-based deep learning models of brain age have defined state-of-the-art performance on large datasets (Cole et al. 2017; Bashyam et al. 2020; Jonsson et al. 2019). However, more importantly, the value of M/EEG-derived brain age models should not be defined in terms of incremental improvement over MRI-based models as M/EEG-based models may enhance MRI-derived information (Engemann et al. 2020) or may be the only option available (Sun et al. 2019).

Our results nicely demonstrate a second critical merit of cross-model and cross-dataset benchmarking. It was sufficient to analyze four different sources of data until we found a perfectly legitimate EEG dataset from an academic research context (CHBP) in which our previously favored modeling techniques developed on the Cam-CAN and the TUAB data did not perform well by default. There may be good reasons for these discrepancies related to the age distribution found in the CHBP data and the fact that multiple different EEG montages were used in that dataset (see results section for detailed discussion in context). But more importantly, we did not anticipate this to happen and would have never learned about it had we confined the scope to previously analyzed datasets. Such discoveries are favored by systematic benchmarks with dataset-independent code implementation, which has the potential to lower the burden threshold for including always more datasets into model development. In the long run, we hope that this effort will stimulate new research leading to more generalizable models.

This brings us to some limitations of this work. Our work has been motivated by the absolute necessity to diversify datasets for development of M/EEG-based measures of brain health. This has led us to analyzing more than 2500 M/EEG recordings and, yet we only included four datasets. Other M/EEG datasets come to mind that would have been potentially relevant. The Human Connectome Project MEG data (Larson-Prior et al. 2013) includes MEG recordings from less than 100 participants, which we deemed insufficient for predictive regression modeling. The OMEGA data resource (Niso et al. 2016) was not accessible at the time of this investigation but would have been a good match for this study. Finally, the LIFE cohort (Loeffler et al. 2015) includes a large number of EEG recordings of participants sampled from the general population yet follows a closed / controlled access scheme. The Healthy-Brain-Network EEG data (Alexander et al. 2017) concerns a developmental cohort. Despite potentially relevant similarities between brain development and aging, age prediction in developmental cohorts would have exceeded the scope of the present study. Even if we had integrated these resources in the present benchmark, this may have only marginally enhanced the diversity covered by the current selection datasets as most public neuroscience datasets come from the wealthiest nations. We hope that this situation will improve as new promising international consortia and efforts emerge that focus on curating large EEG datasets from diverse national and cultural contexts (Ibanez et al. 2021; Ibrahim et al. 2020; "Global Brain Consortium Homepage" n.d.). A second limitation of the present study concerns the depth of validation. To advance our understanding of M/EEG-derived brain age, more systematic comparisons against MRI-derived brain age (Xifra-Porxas et al. 2021) and other measures of mental health and cognitive function are important objectives (Anatürk et al. 2021; Dadi et al. 2021).

In the following we wish to point out a few imminent opportunities for turning the limitations of the present work into future research projects, potentially, enabled by the results and tools brought by the current benchmarks.

Opportunities and suggestions for follow-up research using the benchmark tools

The impact of deeper architectures. An important design decision in deep neural networks is the total depth of the neural network. Here we used previously published architectures designed for EEG-based pathology decoding (Schirrmeister et al. 2017). Future studies could build on top of this benchmark to explore the importance of deep architectures for brain age modeling. Specifically, it would be possible to use methods from neural architecture search, e.g., AutoPyTorch (Zimmer, Lindauer, and Hutter 2021), to design better-performing architectures. Since this benchmark does not only provide access to diverse datasets in an identical file format, but also enables direct comparison to others, it is the optimal starting point for such an optimization while at the same time avoiding overfitting the architecture to a single dataset.

The role of preprocessing. While data cleaning is of major importance for extracting physiologically interpretable biomarkers, predictions from machine learning models tend to be far less affected by noise (Sabbagh et al. 2020). On the other hand, artifacts and noise may inform predictions, potentially reducing biological specificity. Future studies could benefit from this benchmark to quantify the role of artifact signals for brain age predictions and develop de-confounding strategies (Du et al. 2021; Mehrabi et al. 2021; Lu, Schölkopf, and Hernández-Lobato 2018; Bica, Alaa, and Van Der Schaar 2020).

Eyes-open versus eyes-closed. Some of the datasets analyzed in this benchmark contain resting-state signals under different conditions. In the lack of strong a-priori hypotheses, here we simply pooled both conditions. It is currently unclear whether the relationship between eyes-closed versus eyes-open resting-state may contain valuable information about brain aging. It is imaginable, however, that signals induced by transient visual deprivation may reveal levels of vigilance (Wong, DeYoung, and Liu 2016), which in turn may be altered by neuropsychiatric conditions (Hegerl et al. 2012). Future work could benefit from the benchmark to investigate the importance of eyes-closed versus eyes-open resting-state for brain age modeling.

Model averaging. Good prediction performance defines practically useful machine learning models. In many instances, combining prediction models using model averaging approaches can improve the prediction performance (O'Connor et al. 2021; Dadi et al. 2021; Varoquaux et al. 2017). This could also be a practical way of combining the benchmarks into a single model for subsequent generalization testing. Here we demonstrated a very basic form of bagging in which model predictions were combined by their arithmetic means, which did not lead to clear improvements, however. Future studies could use this benchmark to investigate more sophisticated model averaging techniques that combine predictions through supervised learning.

Model inspection. Prediction performance and interpretability can stand in tension with another and are often addressed separately. The interpretability of machine learning models is essential for clinical impact (Rudin 2019; Ghassemi, Oakden-Rayner, and Beam 2021) and can be approached by either analyzing the role of input features for prediction or by analyzing the predictions themselves. In his benchmark we did not cover methods for explaining the role of variable importance for model predictions such as explainers (Biecek 2018; Baniecki et al. 2020) but, instead, only provided basic comparisons between model predictions at the individual-sample level. Applying common explainer methods e.g. SHAP values or permutation importance was not straight-forward in our setting as the benchmarks were based on different types of input data (tabular data for handcrafted benchmark, entire covariance matrices for filterbank models, entire EEG data for deep learning models), preventing apples-to-apples comparisons based on the same set of explainers. Future work could investigate the relative importance of M/EEG signals or features for brain age modeling through a more targeted and systematic analysis using explainers most adequate to each benchmark with subsequent qualitative comparisons. This may also lead to a more general framework for comparing models based on diverse types of input data.

Exploring the link between modalities and cognitive or clinical scores. A second approach to model interpretation is analyzing the correlates of model predictions This study established the tools and methods for basic benchmarks on prediction performance. However, to build useful brain age models, it is essential to validate brain-age predictions to cognitive function, measures of health or clinical endpoints (Dadi et al. 2021; Cole et al. 2018; Liem et al. 2017). To further establish the relative merit of M/EEG over MRI, comparisons between the modalities are essential (Engemann et al. 2020). Unfortunately, direct comparisons between MEG and EEG are not possible at this point as this would require recordings with both modalities in the same subjects. As an intermediate step, it would be worthwhile to use the present benchmarks for developing a model enabling generalization from MEG to EEG and vice versa. In this context, it will be worthwhile to compare against other types of normative approaches for M/EEG beyond brain age (Li et al. 2022). Fortunately, most of the datasets covered in this benchmark include MRI data, social details and psychometric scores next to the M/EEG data. Although these measures are not harmonized across datasets, they still provide a wealth of opportunities for within-dataset validation of brain age measures. Moreover, the benchmarks here could be compared on new emerging datasets and clinical studies.

Conclusion

Computational benchmarks across M/EEG datasets and machine learning methods bear the potential to enhance applications of machine learning in clinical neuroscience in several ways. Standardization of data formats, software and analysis pipelines are important factors for the scalability of predictive modeling of M/EEG. For stimulating the development of more generalizable machine learning models it is crucial that a critical mass of M/EEG datasets be analyzed by the international community. As the diversity of the datasets increases, generalization gaps will manifest themselves, calling for computation methods for closing these gaps. The implied learning process may eventually lead to developing more widely applicable M/EEG-based biomarkers that are clinically robust across a wide range of sociocultural contexts, clinical populations, recording sites and measurement techniques. We hope that benchmarks, tools and resources resulting from this study will facilitate investigating open scientific questions related to learning biomarkers of brain health on an ever-growing number of M/EEG datasets from increasingly diverse real-world contexts.

Declaration of conflicts of interest

D.E. is a full-time employee of F. Hoffmann-La Roche Ltd.
H.B. receives graduate funding support from InteraXon Inc.

Credit authorship contribution statement

Denis A. Engemann: Conceptualization, Data curation, Software, Formal analysis, Supervision, Funding acquisition, Validation, Investigation, Visualization, Methodology, Project administration, Writing – original draft, Writing – review & editing. **Apolline Mellot:** Data curation, Software, Investigation, Writing – review & editing. **Richard Höchenberger:** Software. **Hubert Banville:** Data curation, Software, Investigation, Writing – review & editing. **David Sabbagh:** Software, Writing – review & editing. **Lukas Gemein:** Data curation, Software, Investigation, Writing – review & editing. **Tonio Ball:** Writing – review & editing. **Alexandre Gramfort:** Conceptualization, Data curation, Software, Formal analysis, Supervision, Funding acquisition, Methodology, Writing – review & editing.

Acknowledgements

We would like to thank Pedro Valdés-Sosa and Jorge Bosch-Bayard and their team for the support with the Cuban Human Brain Project data and for the stimulating discussion. We would also like to thank Joseph Picone for the support with Temple University Hospital data. We would like to thank Anahit Babyan, Vadim Nikulin and Arno Villringer for the support with the LEMON dataset and the stimulating scientific discussions. We would like to thank Rik Henson and Ethan Knights for providing the Cam-CAN datasets in BIDS format. We also thank Ana Radanovic and the other participants of the IDESSAI 2021 Inria-DFKI summer school for feedback and testing on earlier versions of the code and analysis developed in this project. This work was supported by the ANR BrAIN AI chair (ANR-20-CHIA-0016) and ANR AI-Cog (ANR-20-IADJ-0002) as well as by the Freiburg Graduate School of Robotics.

References

Al Zoubi, Obada, Wong, Chung Ki, Kuplicki, Rayus T., Yeh, Hung-Wen, Mayeli, Ahmad, Refai, Hazem, Paulus, Martin, Bodurka, Jerzy, 2018. Predicting Age From Brain EEG Signals—A Machine Learning Approach. Frontiers in Aging Neuroscience 10, 184.

Alexander, Lindsay M., Escalera, Jasmine, Ai, Lei, Andreotti, Charissa, Febre, Karina, Mangone, Alexander, Vega-Potler, Natan, et al., 2017. An Open Resource for Transdiagnostic Research in Pediatric Mental Health and Learning Disorders. Scientific Data 4 (December), 170181.

Anatürk, Melis, Kaufmann, Tobias, Cole, James H., Suri, Sana, Griffanti, Ludovica, Zsoldos, Enikő, Filippini, Nicola, et al., 2021. Prediction of Brain Age and Cognitive Age: Quantifying Brain and Cognitive Maintenance in Aging. Human Brain Mapping 42 (6), 1626–1640.

Ang, Kai Keng, Chin, Zheng Yang, Zhang, Haihong, Guan, Cuntai, 2008. Filter Bank Common Spatial Pattern (FBCSP) in Brain-Computer Interface. In: 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence, pp. 2390–2397.

Appelhoff, Stefan, Sanderson, Matthew, Brooks, Teon L., Vliet, Marijn van, Quentin, Romain, Holdgraf, Chris, Chaumon, Maximilien, et al., 2019. MNE-BIDS: Organizing Electrophysiological Data into the BIDS Format and Facilitating Their Analysis. The Journal of Open Source Software 4 (44). https://pure.mpg.de/rest/items/item_3192645/component/file_3192646/content.

Arnold, Jeffrey B., 2017. Ggthemes: Extra Themes, Scales and Geoms for ‘Ggplot2. R Package Version 3 (0).

Babayan, Anahit, Erbey, Miray, Kumral, Deniz, Reinelt, Janis D., Reiter, Andrea M.F., Röbbig, Josefin, Schaare, H.Lina, et al., 2019. A Mind-Brain-Body Dataset of MRI, EEG, Cognition, Emotion, and Peripheral Physiology in Young and Old Adults. Scientific Data 6 (February), 180308.

Babiloni, Claudio, Binetti, Giuliano, Cassarino, Andrea, Forno, Gloria Dal, Del Percio, Claudio, Ferreri, Florinda, Ferri, Raffaele, et al., 2006. Sources of Cortical Rhythms in Adults during Physiological Aging: A Multicentric EEG Study. Human Brain Mapping 27 (2), 162–172.

Baniecki, Hubert, Kretowicz, Wojciech, Piatyszek, Piotr, Wisniewski, Jakub, Biecek, Przemyslaw, 2020. Dalex: Responsible Machine Learning with Interactive Explainability and Fairness in Python. ArXiv [Cs.LG]. arXiv http://arxiv.org/abs/2012.14406 .

Banville, Hubert, Chehab, Omar, Hyvarinen, Aapo, Engemann, Denis, Gramfort, Alexandre, 2020. Uncovering the Structure of Clinical EEG Signals with Self-Supervised Learning. Journal of Neural Engineering doi:10.1088/1741-2552/abca18, November.

Banville, Hubert, Wood, Sean U.N., Aimone, ChrisDenis-Alexander Engemann, and Alexandre Gramfort, 2021. “Robust Learning from Corrupted EEG with Dynamic Spatial Filtering”. ArXiv [Cs.LG]. arXiv http://arxiv.org/abs/2105.12916 .

Bao, Pinglei, She, Liang, McGill, Mason, Tsao, Doris Y., 2020. A Map of Object Space in Primate Inferotemporal Cortex. Nature 583 (7814), 103–108.

Barachant, Alexandre, Bonnet, Stéphane, Congedo, Marco, Jutten, Christian, 2012. Multiclass Brain-Computer Interface Classification by Riemannian Geometry. IEEE Transactions on Bio-Medical Engineering 59 (4), 920–928.

Bashyam, Vishnu M., Erus, Guray, Doshi, Jimit, Habes, Mohamad, Nasrallah, Ilya, Truelove-Hill, Monica, Srinivasan, Dhivya, et al., 2020. MRI Signatures of Brain Age and Disease over the Lifespan Based on a Deep Brain Network and 14 468 Individuals Worldwide. Brain: A Journal of Neurology 143 (7), 2312–2324.

Bica, Ioana, Alaa, Ahmed, Schaar, Mihaela Van Der, 2020. Time Series Deconfounder: Estimating Treatment Effects over Time in the Presence of Hidden Confounders. In: *Proceedings of the 37th International Conference on Machine Learning*, edited by Hal Daumé Iii and Aarti Singh, 119. Proceedings of Machine Learning ResearchPMLR, pp. 884–895.

Biecek, Przemyslaw., 2018. DALEX: Explainers for Complex Predictive Models. ArXiv [Stat.ML]. arXiv http://arxiv.org/abs/1806.08915 .

Bigdely-Shamlo, Nima, Mullen, Tim, Kothe, Christian, Su, Kyung-Min, Robbins, Kay A., 2015. The PREP Pipeline: Standardized Preprocessing for Large-Scale EEG Analysis. Frontiers in Neuroinformatics 9 (June), 16.

Bland, J.M., Altman, D.G., 1999. Measuring Agreement in Method Comparison Studies. Statistical Methods in Medical Research 8 (2), 135–160.

Bosch-Bayard, Jorge, Galan, Lidice, Vazquez, Eduardo Aubert, Virues Alba, Trinidad, Valdes-Sosa, Pedro A., 2020. Resting State Healthy EEG: The First Wave of the Cuban Normative Database. Frontiers in Neuroscience 14 (December), 555119.

Breiman, Leo., 1996. Bagging Predictors. Machine Learning 24 (2), 123–140.

Breiman, Leo., 2001. Random Forests. Machine Learning 45 (1), 5–32.

Buitinck, Lars, Louppe, Gilles, Blondel, Mathieu, Pedregosa, Fabian, Mueller, Andreas, Grisel, Olivier, Niculae, Vlad, et al., 2013. API Design for Machine Learning Software: Experiences from the Scikit-Learn Project. ArXiv [Cs.LG]. arXiv http://arxiv.org/abs/1309.0238 .

Bycroft, Clare, Freeman, Colin, Petkova, Desislava, Band, Gavin, Elliott, Lloyd T., Sharp, Kevin, Motyer, Allan, et al., 2018. The UK Biobank Resource with Deep Phenotyping and Genomic Data. Nature 562 (7726), 203–209.

Cabeza, Roberto, Anderson, Nicole D., Locantore, Jill K., McIntosh, Anthony R., 2002. Aging Gracefully: Compensatory Brain Activity in High-Performing Older Adults. NeuroImage 17 (3), 1394–1402.

Chambon, Stanislas, Galtier, Mathieu N., Arnal, Pierrick J., Wainrib, Gilles, Gramfort, Alexandre, 2018. A Deep Learning Architecture for Temporal Sleep Stage Classification Using Multivariate and Multimodal Time Series. IEEE Transactions on Neural Systems and Rehabilitation Engineering: A Publication of the IEEE Engineering in Medicine and Biology Society 26 (4), 758–769.

Cheveigné, Alain de, Parra, Lucas C., 2014. Joint Decorrelation, a Versatile Tool for Multichannel Data Analysis. NeuroImage 98 (September), 487–505.

Choy, Tricia, Baker, Elizabeth, Stavropoulos, Katherine, 2021. Systemic Racism in EEG Research: Considerations and Potential Solutions. Affective Science doi:10.1007/s42761-021-00050-0, May..

Cole, J.H., Ritchie, S.J., Bastin, M.E., Valdés Hernández, M.C., Muñoz Maniega, S., Royle, N., Corley, J., et al., 2018. Brain Age Predicts Mortality. Molecular Psychiatry 23 (5), 1385–1392.

Cole, James H., 2020. Multimodality Neuroimaging Brain-Age in UK Biobank: Relationship to Biomedical, Lifestyle, and Cognitive Factors. Neurobiology of Aging 92 (August), 34–42.

Cole, James H., Franke, Katja, 2017. Predicting Age Using Neuroimaging: Innovative Brain Ageing Biomarkers. Trends in Neurosciences 40 (12), 681–690.

edited by Cole, James H., Franke, Katja, Cherbuin, Nicolas, 2019. Quantification of the Biological Age of the Brain Using Neuroimaging. In: Moskalev, Alexey (Ed.), Biomarkers of Human Aging. Springer International Publishing, Cham, pp. 293–328 edited by.

Cole, James H., Poudel, Rudra P.K., Tsagkrasoulis, Dimosthenis, Caan, Matthan W.A., Steves, Claire, Spector, Tim D., Montana, Giovanni, 2017. Predicting Brain Age with Deep Learning from Raw Imaging Data Results in a Reliable and Heritable Biomarker. NeuroImage 163 (December), 115–124.

Dadi, Kamalaker, Varoquaux, Gaël, Houenou, Josselin, Bzdok, Danilo, Thirion, Bertrand, Engemann, Denis, 2021. Population Modeling with Machine Learning Can Enhance Measures of Mental Health. GigaScience (10) 10. doi:10.1093/gigascience/giab071.

Dähne, Sven, Meinecke, Frank C., Haufe, Stefan, Höhne, Johannes, Tangermann, Michael, Müller, Klaus-Robert, Nikulin, Vadim V., 2014. SPoC: A Novel Framework for Relating the Amplitude of Neuronal Oscillations to Behaviorally Relevant Parameters. NeuroImage 86 (February), 111–122.

Damoiseaux, J.S., Beckmann, C.F., Sanz Arigita, E.J., Barkhof, F., Scheltens, Ph, Stam, C.J., Smith, S.M., Rombouts, S.A.R.B., 2008. Reduced Resting-State Brain Activity in the 'Default Network' in Normal Aging. Cerebral Cortex 18 (8), 1856–1864.

Denissen, Engemann, De Cock, Costers, Baijot, Laton, Penner, et al., 2021. Brain Age as a Surrogate Marker for Information Processing Speed in Multiple Sclerosis. MedRxiv.

Devarajan, Kavya, Bagyaraj, S., Balasampath, Vinitha, Jyostna, E., Jayasri, K, 2014. EEG-Based Epilepsy Detection and Prediction. IACSIT International Journal of Engineering and Technology 6 (3), 212–216.

Dosenbach, Nico U.F., Nardos, Binyam, Cohen, Alexander L., Fair, Damien A., Power, Jonathan D., Church, Jessica A., Nelson, Steven M., et al., 2010. Prediction of Individual Brain Maturity Using FMRI. Science 329 (5997), 1358–1361.

Driscoll, I., Davatzikos, C., An, Y., Wu, X., Shen, D., Kraut, M., Resnick, S.M., 2009. Longitudinal Pattern of Regional Brain Volume Change Differentiates Normal Aging from MCI. Neurology 72 (22), 1906–1913.

Du, Mengnan, Yang, Fan, Zou, Na, Hu, Xia, 2021. Fairness in Deep Learning: A Computational Perspective. IEEE Intelligent Systems 36 (4), 25–34.

Duncan, L., Shen, H., Gelaye, B., Meijsen, J., Ressler, K., Feldman, M., Peterson, R., Domingue, B., 2019. Analysis of Polygenic Risk Score Usage and Performance in Diverse Human Populations. Nature Communications 10 (1), 3328.

Engemann, Denis A., Kozynets, Oleh, Sabbagh, David, Lemaître, Guillaume, Varoquaux, Gael, Liem, Franziskus, Gramfort, Alexandre, 2020. Combining Magnetoencephalography with Magnetic Resonance Imaging Enhances Learning of Surrogate-Biomarkers. ELife (May) 9. doi:10.7554/eLife.54055.

Engemann, Denis A., Raimondo, Federico, King, Jean-Rémi, Rohaut, Benjamin, Louppe, Gilles, Faugeras, Frédéric, Annen, Jitka, et al., 2018. Robust EEG-Based Cross-Site and Cross-Protocol Classification of States of Consciousness. Brain: A Journal of Neurology 141 (11), 3179–3192.

Esteller, R., Echauz, J., Tcheng, T., Litt, B., Pless, B., 2001. Line Length: An Efficient Feature for Seizure Onset Detection. 2001 Conference Proceedings of the 23rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society 2, 1707–1710 vol.2.

Esteller, R., Vachtsevanos, G., Echauz, J., Litt, B., 2001. A Comparison of Waveform Fractal Dimension Algorithms. IEEE Transactions on Circuits and Systems I: Fundamental Theory and Applications 48 (2), 177–183.

Ewers, Michael, Sperling, Reisa A., Klunk, William E., Weiner, Michael W., Hampel, Harald, 2011. Neuroimaging Markers for the Prediction and Early Diagnosis of Alzheimer's Disease Dementia. Trends in Neurosciences 34 (8), 430–442.

Ferrucci, Luigi, Gonzalez-Freire, Marta, Fabbri, Elisa, Simonsick, Eleanor, Tanaka, Toshiko, Moore, Zenobia, Salimi, Shabnam, Sierra, Felipe, de Cabo, Rafael, 2020. Measuring Biological Aging in Humans: A Quest. Aging Cell 19 (2), e13080.

Fischl, Bruce., 2012. FreeSurfer. NeuroImage 62 (2), 774–781.

Fry, Anna, Littlejohns, Thomas J., Sudlow, Cathie, Doherty, Nicola, Adamska, Ligia, Sprosen, Tim, Collins, Rory, Allen, Naomi E., 2017. Comparison of Sociodemographic and Health-Related Characteristics of UK Biobank Participants with Those of the General Population. American Journal of Epidemiology 186 (9), 1026–1034.

Garcés, Pilar, López-Sanz, David, Maestú, Fernando, Pereda, Ernesto, 2017. Choice of Magnetometers and Gradiometers after Signal Space Separation. Sensors 17 (12). doi:10.3390/s17122926.

Gaubert, Sinead, Raimondo, Federico, Houot, Marion, Corsi, Marie-Constance, Naccache, Lionel, Sitt, Jacobo Diego, Hermann, Bertrand, et al., 2019. EEG Evidence of Compensatory Mechanisms in Preclinical Alzheimer's Disease. Brain: A Journal of Neurology 142 (7), 2096–2112.

Gemein, Lukas A.W., Schirrmeister, Robin T., Chrabąszcz, Patryk, Wilson, Daniel, Boedecker, Joschka, Schulze-Bonhage, Andreas, Hutter, Frank, Ball, Tonio, 2020. Machine-Learning-Based Diagnostics of EEG Pathology. NeuroImage 220 (October), 117021.

Ghassemi, Marzyeh, Oakden-Rayner, Luke, Beam, Andrew L., 2021. The False Hope of Current Approaches to Explainable Artificial Intelligence in Health Care. The Lancet. Digital Health 3 (11), e745–e750.

"Global Brain Consortium Homepage." n.d. Accessed November 30, 2021. https://globalbrainconsortium.org/.

Gonneaud, Julie, Baria, Alex T., Binette, Alexa Pichet, Gordon, Brian A., Chhatwal, Jasmeer P., Cruchaga, Carlos, Jucker, Mathias, et al., 2021. Accelerated Functional Brain Aging in Pre-Clinical Familial Alzheimer's Disease. Nature Communications 12 (1), 1–17.

Gorgolewski, Krzysztof J., Auer, Tibor, Calhoun, Vince D., Craddock, R.Cameron, Das, Samir, Duff, Eugene P., Flandin, Guillaume, et al., 2016. The Brain Imaging Data Structure, a Format for Organizing and Describing Outputs of Neuroimaging Experiments. Scientific Data 3 (June), 160044.

Gramfort, Alexandre, Luessi, Martin, Larson, Eric, Engemann, Denis A., Strohmeier, Daniel, Brodbeck, Christian, Goj, Roman, et al., 2013. MEG and EEG Data Analysis with MNE-Python. Frontiers in Neuroscience 7 (December), 267.

Gramfort, Alexandre, Luessi, Martin, Larson, Eric, Engemann, Denis A., Strohmeier, Daniel, Brodbeck, Christian, Parkkonen, Lauri, Hämäläinen, Matti S., 2014. MNE Software for Processing MEG and EEG Data. NeuroImage 86 (February), 446–460.

Güntekin, Bahar, Aktürk, Tuba, Arakaki, Xianghong, Bonanni, Laura, Del Percio, Claudio, Edelmayer, Rebecca, Farina, Francesca, et al., 2021. Are There Consistent Abnormalities in Event-related EEG Oscillations in Patients with Alzheimer's Disease Compared to Other Diseases Belonging to Dementia? Psychophysiology doi:10.1111/psyp.13934, August.

Harati, A., López, S., Obeid, I., Picone, J., Jacobson, M.P., Tobochnik, S., 2014. The TUH EEG CORPUS: A Big Data Resource for Automated EEG Interpretation. In: 2014 IEEE Signal Processing in Medicine and Biology Symposium (SPMB), pp. 1–5.

Harris, Charles R., Jarrod Millman, K., Walt, Stéfan J.van der, Gommers, Ralf, Virtanen, Pauli, Cournapeau, David, Wieser, Eric, et al., 2020. Array Programming with NumPy. Nature 585 (7825), 357–362.

Hastie, Trevor, Tibshirani, Robert, Friedman, Jerome, Franklin, James, 2005. The Elements of Statistical Learning: Data Mining, Inference and Prediction. The Mathematical Intelligencer 27 (2), 83–85.

He, Tong, Kong, Ru, Holmes, Avram J., Nguyen, Minh, Sabuncu, Mert R., Eickhoff, Simon B., Bzdok, Danilo, Feng, Jiashi, Thomas Yeo, B.T., 2020. Deep Neural Networks and Kernel Regression Achieve Comparable Accuracies for Functional Connectivity Prediction of Behavior and Demographics. NeuroImage 206 (February), 116276.

Hegerl, Ulrich, Wilk, Kathrin, Olbrich, Sebastian, Schoenknecht, Peter, Sander, Christian, 2012. Hyperstable Regulation of Vigilance in Patients with Major Depressive Disorder. The World Journal of Biological Psychiatry: The Official Journal of the World Federation of Societies of Biological Psychiatry 13 (6), 436–446.

Henrich, J., Heine, S., 2010. The Weirdest People in the World? Behavioral and Brain {…}.

Hernandez-Gonzalez, Gertrudis, Bringas-Vega, Maria L., Galán-Garcia, Lidice, Bosch-Bayard, Jorge, Lorenzo-Ceballos, Yenisleidy, Melie-Garcia, Lester, Valdes-Urrutia, Lourdes, Cobas-Ruiz, Marcia, Valdes-Sosa, Pedro ACuban Human Brain Mapping Project (CHBMP), 2011. Multimodal Quantitative Neuroimaging Databases and Methods: The Cuban Human Brain Mapping Project. Clinical EEG and Neuroscience: Official Journal of the EEG and Clinical Neuroscience Society (ENCS) 42 (3), 149–159.

Ibanez, Agustin, Parra, Mario A., Butler, ChristopherLatin America and the Caribbean Consortium on Dementia (LAC-CD), 2021. "The Latin America and the Caribbean Consortium on Dementia (LAC-CD): From Networking to Research to Implementation Science. Journal of Alzheimer's Disease: JAD 82 (s1), S379–S394.

Inouye, T., Shinosaki, K., Sakamoto, H., Toi, S., Ukai, S., Iyama, A., Katsuda, Y., Hirano, M., 1991. Quantification of EEG Irregularity by Use of the Entropy of the Power Spectrum. Electroencephalography and Clinical Neurophysiology 79 (3), 204–210.

Jas, Mainak, Engemann, Denis A., Bekhti, Yousra, Raimondo, Federico, Gramfort, Alexandre, 2017. Autoreject: Automated Artifact Rejection for MEG and EEG Data. NeuroImage 159 (October), 417–429.

Jas, Mainak, Larson, Eric, Engemann, Denis A., Leppäkangas, Jaakko, Taulu, Samu, Hämäläinen, Matti, Gramfort, Alexandre, 2018. A Reproducible MEG/EEG Group Study With the MNE Software: Recommendations, Quality Assessments, and Good Practices. Frontiers in Neuroscience 12, 530.

Jayaram, Vinay, Barachant, Alexandre, 2018. MOABB: Trustworthy Algorithm Benchmarking for BCIs. Journal of Neural Engineering 15 (6), 066011.

Jonsson, B.A., Bjornsdottir, G., Thorgeirsson, T.E., Ellingsen, L.M., Bragi Walters, G., Gudbjartsson, D.F., Stefansson, H., Stefansson, K., Ulfarsson, M.O., 2019. Brain Age Prediction Using Deep Learning Uncovers Associated Sequence Variants. Nature Communications 10 (1), 5409.

Kernbach, Julius M., Thomas Yeo, B.T., Smallwood, Jonathan, Margulies, Daniel S., Schotten, Michel Thiebaut de, Walter, Henrik, Sabuncu, Mert R., et al., 2018. Subspecialization within Default Mode Nodes Characterized in 10,000 UK Biobank Participants. Proceedings of the National Academy of Sciences of the United States of America 115 (48), 12295–12300.

Khan, Sheraz, Hashmi, Javeria A., Mamashli, Fahimeh, Michmizos, Konstantinos, Kitzbichler, Manfred G., Bharadwaj, Hari, Bekhti, Yousra, et al., 2018. Maturation Trajectories of Cortical Resting-State Networks Depend on the Mediating Frequency Band. NeuroImage 174 (July), 57–68.

Kietzmann, Tim C., McClure, Patrick, Kriegeskorte, Nikolaus, 2019. Deep Neural Networks in Computational Neuroscience. Oxford Research Encyclopedia of Neuroscience.

King, J. R., L. Gwilliams, C. Holdgraf, and J. Sassenhagen. 2018. "Encoding and Decoding Neuronal Dynamics: Methodological Framework to Uncover the Algorithms of Cognition." https://hal.archives-ouvertes.fr/hal-01848442/.

King, J-R, Dehaene, S., 2014. Characterizing the Dynamics of Mental Representations: The Temporal Generalization Method. Trends in Cognitive Sciences 18 (4), 203–210.

King, Kevin S., Kozlitina, Julia, Rosenberg, Roger N., Peshock, Ronald M., McColl, Roderick W., Garcia, Christine K., 2014. Effect of Leukocyte Telomere Length on Total and Regional Brain Volumes in a Large Population-Based Cohort. JAMA Neurology 71 (10), 1247–1254.

Kostas, Demetres, Rudzicz, Frank, 2020. Thinker Invariance: Enabling Deep Neural Networks for BCI across More People. Journal of Neural Engineering 17 (5), 056008.

Larson-Prior, L.J., Oostenveld, R., Penna, S.Della, Michalareas, G., Prior, F., Babajani-Feremi, A., Schoffelen, J-M, et al., 2013. Adding Dynamics to the Human Connectome Project with MEG. NeuroImage 80 (October), 190–201.

edited by LeCun, Yann, Haffner, Patrick, Bottou, Léon, Bengio, Yoshua, 1999. Object Recognition with Gradient-Based Learning. In: Forsyth, David A., Mundy, Joseph L., Gesú, Vito di, Cipolla, Roberto (Eds.), Shape, Contour and Grouping in Computer Vision. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 319–345 edited by.

Leeuwen, K.G.van, Sun, H., Tabaeizadeh, M., Struck, A.F., van Putten, M.J.A.M., Westover, M.B., 2019. Detecting Abnormal Electroencephalograms Using Deep Convolutional Networks. Clinical Neurophysiology: Official Journal of the International Federation of Clinical Neurophysiology 130 (1), 77–84.

Leonelli, S. 2016. "Data-Centric Biology." https://www.degruyter.com/document/doi/10.7208/9780226416502/html.

Li, Min, Wang, Ying, Lopez-Naranjo, Carlos, Hu, Shiang, Reyes, Ronaldo César García, Paz-Linares, Deirel, Areces-Gonzalez, Ariosky, et al., 2022. Harmonized-Multinational QEEG Norms (HarMNqEEG). NeuroImage 256 (August), 119190.

Liang, Hualou, Zhang, Fengqing, Niu, Xin, 2019. Investigating Systematic Bias in Brain Age Estimation with Application to Post-Traumatic Stress Disorders. Human Brain Mapping 40 (11), 3143–3152.

Liem, Franziskus, Varoquaux, Gaël, Kynast, Jana, Beyer, Frauke, Masouleh, Shahrzad Kharabian, Huntenburg, Julia M., Lampe, Leonie, et al., 2017. Predicting Brain-Age from Multimodal Imaging Data Captures Cognitive Impairment. NeuroImage 148 (March), 179–188.

Loeffler, Markus, Engel, Christoph, Ahnert, Peter, Alfermann, Dorothee, Arelin, Katrin, Baber, Ronny, Beutner, Frank, et al., 2015. The LIFE-Adult-Study: Objectives and Design of a Population-Based Cohort Study with 10,000 Deeply Phenotyped Adults in Germany. BMC Public Health 15 (July), 691.

Lu, Chaochao, Schölkopf, Bernhard, Hernández-Lobato, José Miguel, 2018. Deconfounding Reinforcement Learning in Observational Settings. ArXiv [Cs.LG]. arXiv http://arxiv.org/abs/1812.10576 .

Mather, Karen Anne, Jorm, Anthony Francis, Parslow, Ruth Adeline, Christensen, Helen, 2011. Is Telomere Length a Biomarker of Aging? A Review. The Journals of Gerontology. Series A, Biological Sciences and Medical Sciences 66 (2), 202–213.

McKinney, WesOthers, 2011. Pandas: A Foundational Python Library for Data Analysis and Statistics. Python for High Performance and Scientific Computing 14 (9), 1–9.

Mehrabi, Ninareh, Morstatter, Fred, Saxena, Nripsuta, Lerman, Kristina, Galstyan, Aram, 2021. A Survey on Bias and Fairness in Machine Learning. ACM Comput. Surv. 115 (6), 1–35 54.

Möller, Sören, Debrabant, Birgit, Halekoh, Ulrich, Petersen, Andreas Kristian, Gerke, Oke, 2021. An Extension of the Bland-Altman Plot for Analyzing the Agreement of More than Two Raters. Diagnostics (Basel, Switzerland) (1) 11. doi:10.3390/diagnostics11010054.

Niso, Guiomar, Gorgolewski, Krzysztof J., Bock, Elizabeth, Brooks, Teon L., Flandin, Guillaume, Gramfort, Alexandre, Henson, Richard N., et al., 2018. MEG-BIDS, the Brain Imaging Data Structure Extended to Magnetoencephalography. Scientific Data 5 (June), 180110.

Niso, Guiomar, Rogers, Christine, Moreau, Jeremy T., Chen, Li-Yuan, Madjar, Cecile, Das, Samir, Bock, Elizabeth, et al., 2016. OMEGA: The Open MEG Archive. NeuroImage 124 (Pt B), 1182–1187.

Nunez, Paul, Srinivasan, Ramesh, 2006. Electric Fields of the Brain: The Neurophysics of EEG. Oxford University Press.

Obeid, Iyad, Picone, Joseph, 2016. The Temple University Hospital EEG Data Corpus. Frontiers in Neuroscience 10 (May), 196.

O'Connor, David, Lake, Evelyn M.R., Scheinost, Dustin, Todd Constable, R., 2021. Resample Aggregating Improves the Generalizability of Connectome Predictive Modeling. NeuroImage 236 (August), 118044.

Oostenveld, R., Praamstra, P., 2001. The Five Percent Electrode System for High-Resolution EEG and ERP Measurements. Clinical Neurophysiology: Official Journal of the International Federation of Clinical Neurophysiology 112 (4), 713–719.

Päivinen, Niina, Lammi, Seppo, Pitkänen, Asla, Nissinen, Jari, Penttonen, Markku, Grönfors, Tapio, 2005. Epileptic Seizure Detection: A Nonlinear Viewpoint. Computer Methods and Programs in Biomedicine 79 (2), 151–159.

Paszke, Adam, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, et al. 2019. "PyTorch: An Imperative Style, High-Performance Deep Learning Library." In *Advances in Neural Information Processing Systems*, edited by H. Wallach, H. Larochelle, A. Beygelzimer, F. d\textquotesingle Alché-Buc, E. Fox, and R. Garnett. Vol. 32. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2019/file/bdbca288fee7f92f2bfa9f701272774 0-Paper.pdf.

Pedersen, Thomas Lin, 2019. Patchwork: The Composer of Plots. R Package Version 1 (0).

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., et al., 2011. Scikit-Learn: Machine Learning in {P}ython. Journal of Machine Learning Research: JMLR 12, 2825–2830.

Pernet, Cyril R., Appelhoff, Stefan, Gorgolewski, Krzysztof J., Flandin, Guillaume, Phillips, Christophe, Delorme, Arnaud, Oostenveld, Robert, 2019. EEG-BIDS, an Extension to the Brain Imaging Data Structure for Electroencephalography. Scientific Data 6 (1), 103.

Perslev, Mathias, Darkner, Sune, Kempfner, Lykke, Nikolic, Miki, Jennum, Poul Jørgen, Igel, Christian, 2021. U-Sleep: Resilient High-Frequency Sleep Staging. NPJ Digital Medicine 4 (1), 72.

Poldrack, Russell A., Huckins, Grace, Varoquaux, Gael, 2020. Establishment of Best Practices for Evidence for Prediction: A Review. JAMA Psychiatry (Chicago, Ill.) 77 (5), 534–540.

Raffel, Joel, Cole, James, Record, Chris, Sridharan, Sujata, Sharp, David, Nicholas, Richard, 2017. Brain Age: A Novel Approach to Quantify the Impact of Multiple Sclerosis on the Brain (P1.371). Neurology 88 (16 Supplement). https://n.neurology.org/content/88/16_Supplement/P1.371.short.

Richman, J.S., Moorman, J.R., 2000. Physiological Time-Series Analysis Using Approximate Entropy and Sample Entropy. American Journal of Physiology. Heart and Circulatory Physiology 278 (6), H2039–H2049.

Roberts, S.J., Penny, W., Rezek, I., 1999. Temporal and Spatial Complexity Measures for Electroencephalogram Based Brain-Computer Interfacing. Medical & Biological Engineering & Computing 37 (1), 93–98.

Rodrigues, Pedro Luiz Coelho, Jutten, Christian, Congedo, Marco, 2019. Riemannian Procrustes Analysis: Transfer Learning for Brain-Computer Interfaces. IEEE Transactions on Bio-Medical Engineering 66 (8), 2390–2401.

Roy, Yannick, Banville, Hubert, Albuquerque, Isabela, Gramfort, Alexandre, Falk, Tiago H., Faubert, Jocelyn, 2019. Deep Learning-Based Electroencephalography Analysis: A Systematic Review. Journal of Neural Engineering 16 (5), 051001.

Rudin, Cynthia., 2019. Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead. Nature Machine Intelligence 1 (5), 206–215.

Sabbagh, D., Ablin, P., Varoquaux, G., Gramfort, A., 2019. Manifold-Regression to Predict from MEG/EEG Brain Signals without Source Modeling. ArXiv Preprint ArXiv. https://arxiv.org/abs/1906.02687.

Sabbagh, David, Ablin, Pierre, Varoquaux, Gaël, Gramfort, Alexandre, Engemann, Denis A., 2020. Predictive Regression Modeling with MEG/EEG: From Source Power to Signals and Cognitive States. NeuroImage 222 (November), 116893.

Scahill, Rachael I., Frost, Chris, Jenkins, Rhian, Whitwell, Jennifer L., Rossor, Martin N., Fox, Nick C., 2003. A Longitudinal Study of Brain Volume Changes in Normal Aging Using Serial Registered Magnetic Resonance Imaging. Archives of Neurology 60 (7), 989–994.

Schiratti, J-B, Douget, Jean-Eudes Le, Le Van Quyen, Michel, Essid, Slim, Gramfort, Alexandre, 2018. An Ensemble Learning Approach to Detect Epileptic Seizures from Long Intracranial EEG Recordings. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 856–860.

Schiratti, J-B, Douget, Jean-Eudes Le, Le Van Quyen, Michel, Essid, Slim, Gramfort, Alexandre, 2018. An Ensemble Learning Approach to Detect Epileptic Seizures from Long Intracranial EEG Recordings. 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) doi:10.1109/icassp.2018.8461489.

Schirrmeister, Robin Tibor, Springenberg, Jost Tobias, Fiederer, Lukas Dominique Josef, Glasstetter, Martin, Eggensperger, Katharina, Tangermann, Michael, Hutter, Frank, Burgard, Wolfram, Ball, Tonio, 2017. Deep Learning with Convolutional Neural Networks for EEG Decoding and Visualization. Human Brain Mapping 38 (11), 5391–5420.

Schulz, M.A., Bzdok, D., Haufe, S., Haynes, J.D., Ritter, K., 2022. Performance Reserves in Brain-Imaging-Based Phenotype Prediction. BioRxiv. https://www.biorxiv.org/content/10.1101/2022.02.23.481601.abstract.

Schulz, Marc-Andre, Thomas Yeo, B.T., Vogelstein, Joshua T., Mourao-Miranada, Janaina, Kather, Jakob N., Kording, Konrad, Richards, Blake, Bzdok, Danilo, 2020. Different Scaling of Linear Models and Deep Learning in UKBiobank Brain Images versus Machine-Learning Datasets. Nature Communications doi:10.1038/s41467-020-18037-z.

Schumacher, Julia, Ray, Nicola J., Hamilton, Calum A., Donaghy, Paul C., Firbank, Michael, Roberts, Gemma, Allan, Louise, et al., 2021. Cholinergic White Matter Pathways in Dementia with Lewy Bodies and Alzheimer's Disease. Brain: A Journal of Neurology doi:10.1093/brain/awab372, October.

Shafto, Meredith A., Tyler, Lorraine K., Dixon, Marie, Taylor, Jason R., Rowe, James B., Cusack, Rhodri, Calder, Andrew J., et al., 2014. The Cambridge Centre for Ageing and Neuroscience (Cam-CAN) Study Protocol: A Cross-Sectional, Lifespan, Multidisciplinary Examination of Healthy Cognitive Ageing. BMC Neurology 14 (October), 204.

Ibrahim, Shekh, Aida, Sharifah, Hamzah, Nurfaten, Wahab, Athirah Raihanah Abdul, Abdullah, Jafri Malin, Malim, Nurul Hashimah Ahamed Hassain, Sumari, Putra, Idris, Zamzuri, et al., 2020. Big Brain Data Initiative in Universiti Sains Malaysia: Challenges in Brain Mapping for Malaysia. The Malaysian Journal of Medical Sciences: MJMS 27 (4), 1–8.

Sitt, Jacobo Diego, King, Jean Remi, Karoui, Imen El, Rohaut, Benjamin, Faugeras, Frederic, Gramfort, Alexandre, Cohen, Laurent, Sigman, Mariano, Dehaene, Stanislas, Naccache, Lionel, 2014. Large Scale Screening of Neural Signatures of Consciousness in Patients in a Vegetative or Minimally Conscious State. Brain: A Journal of Neurology 137 (8), 2258–2270.

Smith, Stephen M., Nichols, Thomas E., Vidaurre, Diego, Winkler, Anderson M., Behrens, Timothy E.J., Glasser, Matthew F., Ugurbil, Kamil, Barch, Deanna M., Essen, David C.Van, Miller, Karla L., 2015. A Positive-Negative Mode of Population Covariation Links Brain Connectivity, Demographics and Behavior. Nature Neuroscience 18 (11), 1565–1567.

Smith, Stephen M., Vidaurre, Diego, Alfaro-Almagro, Fidel, Nichols, Thomas E., Miller, Karla L., 2019. Estimation of Brain Age Delta from Brain Imaging. NeuroImage 200 (October), 528–539.

Spiegelhalter, David., 2016. How Old Are You, Really? Communicating Chronic Risk through 'Effective Age' of Your Body and Organs. BMC Medical Informatics and Decision Making 16 (1). doi:10.1186/s12911-016-0342-z.

Stokes, Mark G., Wolff, Michael J., Spaak, Eelke, 2015. Decoding Rich Spatial Information with High Temporal Resolution. Trends in Cognitive Sciences 19 (11), 636–638.

Sun, Haoqi, Paixao, Luis, Oliva, Jefferson T., Goparaju, Balaji, Carvalho, Diego Z., Leeuwen, Kicky G.van, Akeju, Oluwaseun, et al., 2019. Brain Age from the Electroencephalogram of Sleep. Neurobiology of Aging 74 (February), 112–120.

Taulu, Samu, Simola, Juha, Kajola, Matti, 2005. Applications of the Signal Space Separation Method." *Signal Processing*. IEEE Transactions On 53 (9), 3359–3372.

Taylor, Jason R., Williams, Nitin, Cusack, Rhodri, Auer, Tibor, Shafto, Meredith A., Dixon, Marie, Tyler, Lorraine K., Henson, Richard N.Others, 2017. The Cambridge Centre for Ageing and Neuroscience (Cam-CAN) Data Repository: Structural and Functional MRI, MEG, and Cognitive Data from a Cross-Sectional Adult Lifespan Sample. NeuroImage 144, 262–269.

Teixeira, C.A., Direito, B., Feldwisch-Drentrup, H., Valderrama, M., Costa, R.P., Alvarado-Rojas, C., Nikolopoulos, S., et al., 2011. EPILAB: A Software Package for Studies on the Prediction of Epileptic Seizures. Journal of Neuroscience Methods 200 (2), 257–271.

Tibor Schirrmeister, Robin, Gemein, Lukas, Eggensperger, Katharina, Hutter, Frank, Ball, Tonio, 2017. Deep Learning with Convolutional Neural Networks for Decoding and Visualization of EEG Pathology. ArXiv E-Prints AugustarXiv:1708.08012.

Tietz, Marian, T. J. Fan, D. Nouri, and Others. 2017. "Skorch: A Scikit-Learn Compatible Neural Network Library That Wraps PyTorch." July.

Valdes-Sosa, Pedro A., Galan-Garcia, Lidice, Bosch-Bayard, Jorge, Bringas-Vega, Maria L., Aubert-Vazquez, Eduardo, Rodriguez-Gil, Iris, Das, Samir, et al., 2021. The Cuban

Human Brain Mapping Project, a Young and Middle Age Population-Based EEG, MRI, and Cognition Dataset. Scientific Data 8 (1), 45.

Van Essen, David C., Smith, Stephen M., Barch, Deanna M., Behrens, Timothy E.J., Yacoub, Essa, Ugurbil, Kamil, Consortium, WU-Minn HCP, 2013. The WU-Minn Human Connectome Project: An Overview. NeuroImage 80 (October), 62–79.

Varoquaux, Gaël, Reddy Raamana, Pradeep, Engemann, Denis A.Andrés Hoyos-Idrobo, Yannick Schwartz, and Bertrand Thirion, 2017. Assessing and Tuning Brain Decoders: Cross-Validation, Caveats, and Guidelines. NeuroImage 145 (Pt B), 166–179.

Virtanen, Pauli, Gommers, Ralf, Oliphant, Travis E., Haberland, Matt, Reddy, Tyler, Cournapeau, David, Burovski, Evgeni, et al., 2020. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. Nature Methods 17 (3), 261–272.

Völker, Martin, Schirrmeister, Robin T., Fiederer, Lukas D.J., Burgard, Wolfram, Ball, Tonio, 2018. Deep Transfer Learning for Error Decoding from Non-Invasive EEG. In: 2018 6th International Conference on Brain-Computer Interface (BCI), pp. 1–6.

Walhovd, K.B., Fjell, A.M., Brewer, J., McEvoy, L.K., Fennema-Notestine, C., Hagler Jr, D.J., Jennings, R.G., Karow, D., Dale, A.M.Alzheimer's Disease Neuroimaging Initiative, 2010. Combining MR Imaging, Positron-Emission Tomography, and CSF Biomarkers in the Diagnosis and Prognosis of Alzheimer Disease. AJNR. American Journal of Neuroradiology 31 (2), 347–354.

Wickham, Hadley., 2011. Ggplot2. Wiley Interdisciplinary Reviews. Computational Statistics 3 (2), 180–185.

Wong, Chi Wah, DeYoung, Pamela N., Liu, Thomas T., 2016. Differences in the Resting-State FMRI Global Signal Amplitude between the Eyes Open and Eyes Closed States Are Related to Changes in EEG Vigilance. NeuroImage 124 (Pt A), 24–31.

Wrigglesworth, Jo, Yaacob, Nurathifah, Ward, Phillip, Woods, Robyn L., McNeil, John, Storey, Elsdon, Egan, Gary, et al., 2021. Brain-Predicted Age Difference Is Associated with Cognitive Processing in Later-Life. Neurobiology of Aging doi:10.1016/j.neurobiolaging.2021.10.007, October.

Xifra-Porxas, Alba, Ghosh, Arna, Mitsis, Georgios D., Boudrias, Marie-Hélène, 2021. Estimating Brain Age from Structural MRI and MEG Data: Insights from Dimensionality Reduction Techniques. NeuroImage 231 (117822), 117822.

Yamins, Daniel L.K., DiCarlo, James J., 2016. Using Goal-Driven Deep Learning Models to Understand Sensory Cortex. Nature Neuroscience 19 (3), 356–365.

Ye, Elissa, Sun, Haoqi, Leone, Michael J., Paixao, Luis, Thomas, Robert J., Lam, Alice D., Brandon Westover, M., 2020. Association of Sleep Electroencephalography-Based Brain Age Index With Dementia. JAMA Network Open 3 (9), e2017357.

Yger, Florian, Berar, Maxime, Lotte, Fabien, 2017. Riemannian Approaches in Brain-Computer Interfaces: A Review. IEEE Transactions on Neural Systems and Rehabilitation Engineering: A Publication of the IEEE Engineering in Medicine and Biology Society 25 (10), 1753–1762.

Zimmer, Lucas, Lindauer, Marius, Hutter, Frank, 2021. Auto-Pytorch: Multi-Fidelity MetaLearning for Efficient and Robust AutoDL. IEEE Transactions on Pattern Analysis and Machine Intelligence 43 (9), 3079–3090.

Further reading

Schulz, Marc-Andre, Thomas Yeo, B.T., Vogelstein, Joshua T., Mourao-Miranada, Janaina, Kather, Jakob N., Kording, Konrad, Richards, Blake, Bzdok, Danilo, 2020. Different Scaling of Linear Models and Deep Learning in UKBiobank Brain Images versus Machine-Learning Datasets. Nature Communications 11 (1), 4238.