Ilaria Boscolo Galazzo, Federica Cruciani, Lorenza Brusini, Ahmed Salih, Petia Radeva, Silvia Francesca Storti, and Gloria Menegaz

Explainable Artificial Intelligence for Magnetic Resonance Imaging Aging Brainprints

Grounds and challenges

SHUTTERSTOCK.COM/KKSSR

Marked changes occur in the brain during people's lives, and individual rates of aging have revealed pronounced differences, giving rise to subject-specific brainprints that are the signature of the brain. These are shaped by a great variety of factors, both endogenous and exogenous. Accurate predictions of brain age (BA) can be derived from neuroimaging endophenotypes by using machine and deep learning (DL) techniques. Predictive models leading to accurate estimates while revealing which features contribute the most to final predictions are key to unveiling the mechanisms underlying the evolution of brain aging patterns. Explainable artificial intelligence (XAI) methods are emerging as enabling technology in different fields, and biomedicine is no exception. Within this framework, this article examines BA and presents a comprehensive review of recent advances in the exploitation of explainable machine learning (ML)/DL methods, highlighting the main open issues and providing hints for future directions.

Introduction

How old is one's brain? This apparently simple question hides an extremely complex system where endogenous and exogenous variables of different types interplay in a still-unknown manner. In this article, we aim to provide a glimpse of the scene by focusing on a specific case: explaining the impact of neuroimaging-derived endophenotypes on determining the brainprint, that, is the "brain signature" or "fingerprint" of the brain, while exploiting XAI for unveiling the main factors ruling the process. The study of brain aging has recently gained attention in the scientific community since developing accurate biomarkers for BA by relying on neuroimaging data in combination with ad hoc statistical analyses opens new perspectives in different domains, enabling us to disentangle age-related from disease-specific changes and track disease progression at the single-subject level [1].

The prediction model, generally trained on large samples of controls, is fed with candidate endophenotypes, and it outputs the estimated, or predicted, age [2]. A so-called delta, or gap, is then defined, given by the difference between predicted and chronological age [3]. Deltas [hereafter referred to as brain-predicted

Digital Object Identifier 10.1109/MSP.2021.3126573
Date of current version: 24 February 2022

age delta (brain-PAD) [4]] reflect individuals' deviation from the population norm, highlighting accelerated aging (positive deltas) and resilience to aging (negative deltas) [3], thus providing information about brain health. Brain-PAD measures are of value for assessing normal aging and disease, with recent studies revealing patterns of faster aging in several neurological and psychiatric pathologies, even prior to overt disease manifestations [1], [2]. An important example can be found in the context of neurodegenerative conditions, where an initial study by the authors of [5] demonstrated significant differences between brain-PAD scores of controls/stable mild cognitive impairment (MCI) and Alzheimer's disease (AD) patients and a more accurate prediction of conversion to AD when using brain-PAD scores rather than neuropsychological tests.

Different solutions have been proposed for tackling this problem, from the choice of the endophenotypes to the methodologies proposed for predicting BA. In the state of the art (SOA), such methodologies range from classical linear regression to ML models working with single/multimodal data. The advent of publicly available large repositories of heterogeneous data called for new methods to cope with high data dimensionality, with DL being first in line [6]. This made vital the explainability/interpretability of the models' outcomes, especially considering the lack of ground truth that is inherent to BA estimation.

Considering the complexity and interdisciplinary nature of BA, we aim at providing a SOA overview of the different approaches to BA prediction, highlighting their respective strengths and weaknesses in light of new challenges and detailing the most important steps in the BA framework, with a special focus on XAI methods used in the brain aging field so far. The main keywords being *neuroimaging*, *DL*, and *XAI*, this review spans these topics, emphasizing the value that they jointly and respectively bring to the problem at hand, and provides hints for potential research directions. The research was carried out by taking inspiration from Preferred Reporting Items for Systematic Reviews and Meta-Analyses guidelines. More details and a synthesis of the selected articles can be found in the supplementary materials available at http://doi.org/10.1109/MSP.2021.3126573. In doing this, particular emphasis was devoted to the exploitation of XAI. Despite XAI methods being scarcely investigated in this respect to date, they hold great potential as a means to shed light on the process, as in [7] and [8].

Neuroimaging-based phenotypes for BA prediction

In this section, an overview of the most common image-derived endophenotypes (IDPs) used in the current literature as imaging features to predict BA is provided, with a focus on magnetic resonance imaging (MRI)-based features. Neuroimaging techniques represent natural instruments to assess the impact of genetic and environmental factors on brain structure and function. T1-weighted (T1w) MRI images have demonstrated that the brain encounters progressive atrophy across its life span [9]. However, premature anatomical changes have been associated with several conditions, such as neurodegenerative diseases, hypertension, and obesity, flagging the brain-PAD as a means for characterizing one's health [10].

In the current literature, most predictive models for BA rely on T1w images as inputs [4], given their greater availability, reliability, and ease of interpretation. This approach was followed in 14 of the considered papers, while eight other works considered additional modalities. Depending on the granularity and framework, several features are generally extracted from T1w images to be used as predictors in BA models. The easiest solution consists of using raw whole-brain T1w data, avoiding the step of feature engineering. However, given the high dimensionality of these data, this is currently feasible only using DL. Six of the considered papers followed this approach, relying on 3D convolutional neural networks (CNNs) [6]–[8], [11]–[13].

In nine of the chosen articles, in addition or alternatively, T1w images were segmented to derive gray matter (GM)/white matter (WM) tissue probability maps to be used as 3D inputs [6], [8], [13] or after being vectorized [5], [9], [14]–[17]. In addition, two works performed tensor-based morphometry on T1w data to use the derived Jacobian maps as novel endophenotypes [13], [18]. Conversely, in 10 studies, the authors moved from voxel-based to region-based approaches and extracted summary statistics for different regions of interest (ROIs), in particular, cortical thickness, surface area, and volume. Overall, models relying on such features exploit information about the atrophy level, such as changes in tissue volume, cortical thinning, and sulcal widening, to provide accurate BA predictions.

Nevertheless, neuroimaging modalities other than conventional T1w MRI can complement the picture provided by these data and inform on other relevant aspects, such as tissue microstructure and brain functioning. See "Mapping the Brain With Functional and Diffusion Magnetic Resonance Imaging Techniques: From Voxel/Region- Wise Measures to Complex Network-Based Metrics." Concerning structural MRI, T2-weighted (T2w) images represent a viable complement to T1w data. This approach was followed, in particular, in UK Biobank (UKB) data processing [3], [4], [10]. The main features derived from such images are represented by the T1w/T2w surface ratio, averaged across all voxels in a given ROI to provide a simple measure of the myelin content and WM lesion volume. Diffusion-weighted imaging [diffusion MRI (dMRI)] represents an advanced structural technique capturing microstructural features of brain tissues in vivo. Several microstructural indices can be derived by fitting raw dMRI signals to specific representation models. All the articles considered here employed the simple tensor model from which fractional anisotropy (FA) and diffusivity measures, such as mean diffusivity and radial diffusivity, were derived.

Of note, *anisotropy* refers to water molecule movement hindrance, while *diffusivity* measures how water molecules are free to diffuse in a given direction. The UKB also included additional information from the neurite orientation dispersion and density imaging model (NODDI), e.g., the intracellular volume fraction, isotropic volume fraction, and orientation dispersion, while, in one work, indices were derived from the mean apparent propagator (MAP) model, providing different measures, including generalized FA (GFA), non-Gaussianity (NG), orthogonal NG, and parallel NG indices [19]. Starting from these maps, the IDPs generally extracted are represented by the mean values calculated

Recording and first-level analysis

Diffusion magnetic resonance imaging (dMRI) is a variant of structural MRI that provides details about microstructural properties by fitting mathematical models (either analytical or compartmental) to raw data (Figure S1). In addition, a structural connectome can be built by tracking the main diffusion direction in each voxel. Blood oxygenation level-dependent contrast functional MRI (fMRI) is the most common noninvasive approach to record the functional activity of the brain, while arterial spin labeling (ASL) represents a viable alternative enabling the derivation of quantitative cerebral blood flow maps along with comparable functional information. Both techniques can be applied during the execution of a given task or while resting (resting-state fMRI). To derive meaningful patterns, various computational/statistical methods are available (e.g., a general linear model for task-fMRI and related functional localization, and independent component analysis for resting-state fMRI).

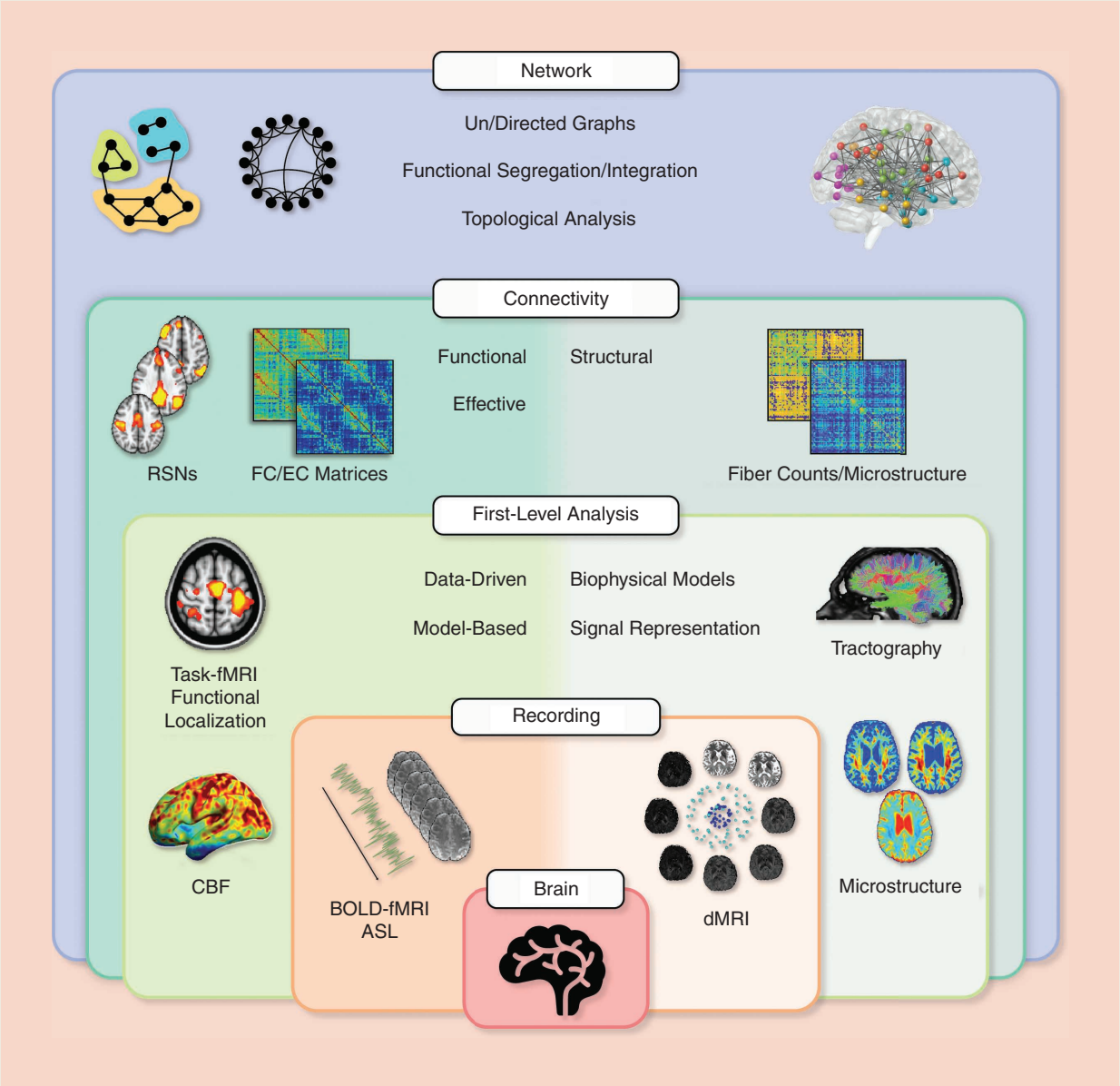

FIGURE S1. An overview of advanced MRI techniques and different feature sets that can be derived from first-level to network analyses. FC: functional connectivity; EC: effective connectivity; fMRI: functional MRI; CBF: cerebral blood flow; BOLD: blood oxygenation level-dependent contrast; ASL: arterial spin labeling; RSNs: resting state networks.

(*Continued*)

Mapping the Brain With Functional and Diffusion Magnetic Resonance Imaging Techniques: From Voxel/Region-Wise Measures to Complex Network-Based Metrics (*Continued*)

Connectivity

Connectivity enables characterizing the interplay among brain regions, either in the form of a backbone of connections (structural connectivity) or functional dependencies [functional connectivity (FC)/effective connectivity (EC)]. From a structural point of view, connectivity is represented by the white matter streamlines linking different brain regions. Such links can be quantified by the normalized number of fibers or summary statistics of microstructural indices collected along fiber bundles. Functional measures capture patterns of statistical dependence among neural elements and, in the case of EC, directed causal effects. While Pearson correlation is generally used for estimating FC, different measures have been proposed for EC, such as dynamic causal modeling and Granger causality.

Network

Network measures can be derived to describe brain structure and function, either separately or jointly. A network consists of a set of nodes (neural elements) and edges (mutual connections). Brain networks can be derived from physiological observations and anatomy, resulting in functional and structural networks. These can then be further examined, with methods of network science extracting measures of segregation, integration, and influence. For more details about the analysis pipelines for dMRI, fMRI and ASL (from preprocessed data to the modality-specific features) please refer to the supplementary material (available in https://doi.org/10.1109/MSP.2021.3126573) with this article.

across WM maps or along different tracts, the latter identified with either tract-based spatial statistics [3], [4], [10], [20], [21] or tractography [4], [10], [19]. Overall, BA models enclosing these features capture age-related changes in the WM microstructure, such as a decrease in anisotropy, which mainly reflects demyelination, axonal degeneration, or neuroinflammation through edema formation.

Besides analyzing brain architecture, important information about functionality can be extracted by relying on functional MRI (fMRI) based on blood oxygenation level-dependent contrast (BOLD) and arterial spin labeling (ASL). IDPs based on BOLD fMRI data were adopted in four articles [20]–[23], alongside those studies employing UKB data [3], [4], [10]. FMRI scans can be acquired either during the execution of a given task (task-fMRI) or while resting (resting-fMRI). Among the works selected here, task-related IDPs were used only within the UKB framework [3], [4], [10] and represented by activation measures in regions derived by group-level maps (i.e., the median and 90th percentile for both the percent signal change and z-statistics). Conversely, several IDPs based on resting-fMRI data have been explored in this context. While measures related to the amplitude of low-frequency fluctuations and regional homogeneity were reported in a single work [21], features describing functional connectivity (FC) patterns were usually employed in such studies.

In more detail, the signals extracted from preprocessed resting-fMRI data were analyzed using different statistical measures, among which Pearson full and partial correlations were the most common [3], [4], [10], [20], [23], reflecting the synchronicity between pairs in regions. Then, adjacency matrices were derived, and the entries of such connectivity matrices were used to feed BA prediction models. Only one study stood out for using covariance as a connectivity measure [22]. In contrast, Rokicki et al. [24] exploited ASL-based regional measures of cerebral blood flow (CBF), which are tightly correlated to neuronal activation and reflect microvascular integrity/functionality, and used them as BA predictors. Finally, two works [17], [23] proposed an attempt to combine MRI with features derived from electrophysiological measures, such as magnetoencephalography (MEG). In this case, the IDPs were constituted by features such as the power spectral density, amplitude envelope correlation, interlayer coupling, source activity (e.g., the signal power), and source connectivity (e.g., the signal covariance).

One of the main issues affecting BA prediction is the unbalance between the number of IDPs and the sample size, which could lead to model overfitting. To overcome this bottleneck, dimensionality reduction is usually applied before feeding IDPs to prediction models. This is generally based on principal component analysis (PCA) [5], [16], [17], canonical correlation analysis (CCA) [17], and independent component analysis (ICA) [10], [14]. These were recently complemented by feature handcrafting, as in [20]. It might be useful to remark that while PCA projects data along the dimensions of maximum variance, CCA maximizes a given similarity measure, most commonly correlation, among the data [17]. When ICA is applied for dimensionality reduction, data are projected in a space where the components are assumed to be non-Gaussian and as independent as possible. Such data projection can provide additional information about a population, reflecting its intrinsic variance and similarity.

Large-scale biomedical databases were used in all the considered contributions, such as the Alzheimer's Disease Neuroimaging Initiative (ADNI), UKB, Cambridge Center for Aging and Neuroscience (Cam-CAN), and Information Extraction From Images (IXI) (please refer to Supplementary Table 1 in the supplementary material available on *Xplore* with this article for more details). No other endophenotypes were used in the papers that were considered. Indeed, traits having a clear link with the aging process (e.g., gender, education, blood pressure, and clinical variables) were used only as covariates and for conducting association studies after BA was predicted, as we will detail in the "BA Prediction Modeling: From Statistical Methods to DL" and "Association Studies With BA Findings" sections.

Overall, the key points concerning IDPs and BA can be summarized as follows:

1) Structural T1w data are largely adopted in BA studies, providing morphometric information ranging from basic tissue probability maps to more complex measures quantifying cortical thickness, surface area, volume, and sulcal width.
2) DMRI, BOLD fMRI, and ASL represent more advanced techniques to complement conventional T1w-based IDPs. They facilitate deriving a wide number of features describing microstructural, functional, and hemodynamic brain patterns, which are currently pursued in the BA field, with promising results.
3) The use of endophenotypes of different natures, for example, based on electrophysiological measures, is still limited, while no studies have exploited other traits with clear links to the aging process.
4) The possible unbalance between the number of IDPs and the sample size is an important issue that could bias BA predictions and needs to be accounted for (e.g., through dimensionality reduction approaches).

BA prediction modeling: From statistical methods to DL

In this section, a detailed overview of the BA estimation framework is provided, first introducing the more conventional statistical/ML methods and related findings, then illustrating the DL-based approaches exploited in the BA field so far. A growing body of research is applying several supervised, linear, and nonlinear techniques to the problem of BA prediction. Multivariate analyses facilitate accurately detecting even subtle deviations from expected age-related brain patterns in individuals [1]. The "BA gap estimation" method [16], developed in 2010, was the first to quantify the acceleration/deceleration of individual brain aging from T1w data, revealing promising results in different studies. Several alternatives have been recently devised, going beyond conventional statistical/ML approaches and leveraging the considerable promises of DL methods, as we will summarize in the following.

Statistical methods/ML

The general process for predicting BA, illustrated in Figure 1, relies on the identification of a large sample of healthy subjects without neuropsychiatric, neurological, and other health disorders, representing the "training set." Once their IDPs are extracted, the following steps are generally taken:

1) *Definition of the prediction model*: From a design perspective, most BA studies use large training sets of subjects within a supervised learning framework to build an age prediction model, having brain IDPs as independent (predictor) variables along with chronological age as the dependent (outcome) variable. Simple linear regression and its extension to accommodate multiple predictors [multiple linear regression (MLR)] have been proposed as simple though effective methods to model the variables' statistical relationship. Given n subjects and p brain features, MLR can be formulated as

$$Y = X\beta + \varepsilon = \beta_0 + X_1 \cdot \beta_1 + X_2 \cdot \beta_2 + \cdots + X_p \cdot \beta_p + \varepsilon, \quad (1)$$

where $Y \in \mathbb{R}^n$ is the vector of the response variable (chronological age), $X \in \mathbb{R}^{n \times p}$ is the independent variable matrix (brain features), $\beta \in \mathbb{R}^p$ is a weight vector (unknown coefficients to be estimated), and ε is a vector of random errors.

However, given the high dimensionality of the neuroimaging features currently available, conventional ordinary least-squares (OLS) regression methods might be inappropriate, leading to the overfitting of the model when a numerosity-matched cohort of subjects is not available. Hence, valid alternatives are represented by more advanced multivariate ML methods that are able to cope with large numbers of features and big data repositories. BA models based on eight different ML methods have been investigated in the papers included in this review, namely, regularized regression models [Ridge, the least absolute shrinkage and selection operator (LASSO), and Elastic Net], Gaussian process regression (GPR), support vector regression (SVR), relevance vector regression (RVR), random forest (RF) and gradient boosting regression (GBR). Their main characteristics are briefly summarized in the supplementary material (available in https://doi.org/10.1109/MSP.2021.3126573) with this article.

2) *Training and validation*: At this stage, IDPs and participants' chronological ages from the training set are fed into the chosen ML model. To validate a BA model, most studies employ a cross-validation approach in which a proportion of the samples from the entire training group is left out (typically ranging between 10 and 20%), while the remaining largest portion is used to train the BA model. This is then applied to the left-out group to predict individual ages. This operation is performed until the whole set of disjoint partitions has been explored. The model's performance is evaluated relying on predefined measures, typically the mean absolute error (MAE), root-mean-square error, and Pearson correlation coefficient (r-value) between the estimated BA and chronological age. While these measures are largely used to assess the accuracy of models, they should be interpreted with caution, especially when comparing the results across studies, as they are affected by several factors, including the age range of the sample, which could lead to changes in performance.
3) *Application of the prediction model to an independent test set*: Usually, the best model is retained for testing on an unseen set of samples (the testing set), consisting of healthy and/or diseased subjects and generating individual predictions. This operation enables further validating a BA model and proving its generalizability across several samples, possibly coming from different sources and databases. Finally, the brain-PAD is extracted as single representative metric given by the difference between the predicted and chronological ages [25].

Key results from BA studies using statistical/ML methods

All the relevant studies included here regarding a statistical/ML perspective (17 papers plus two studies comparing DL and ML predictions) followed the general path described in the preceding.

The choice of the regression algorithm varies considerably in the SOA, with RVR being highly popular as part of the well-known "BA gap estimation" framework [16]. However, in this distilled selection of papers, all the previously mentioned regression methods are equally represented, with no clear prevalence of one approach among the others. While a few authors still rely on simple OLS [3], [10], [22], the current literature is moving toward more sophisticated ML approaches to take advantage of their properties (e.g., robustness to multicollinearity for Ridge regression and robustness to outliers and high generalization for RVR). Two studies moved a step further and proposed a stacked model combining Ridge/GPR with an RF in a multimodal approach [17], [23], with promising results, especially in the presence of missing data.

Finally, a single paper tackled this problem by going beyond ML and relying on a statistically grounded workflow integrating a functional data analysis framework with quantile regression [18]. In particular, this method provided not only a point estimate but also a prediction interval, which could further improve the utility of these BA measures. In terms of performance, a quantitative comparison is hampered by the use of different data sets, imaging features, and age ranges across studies. However, following the findings reported by several researchers, the choice of the ML algorithm does not appear to have a strong bearing on model accuracy. Indeed, comparative studies demonstrated similar performance across methods when applied to the same data, as observed for Ridge, SVR and GPR [21], Elastic Net, GPR and GBR [26] or Ridge, Elastic Net, LASSO, and SVR and RVR

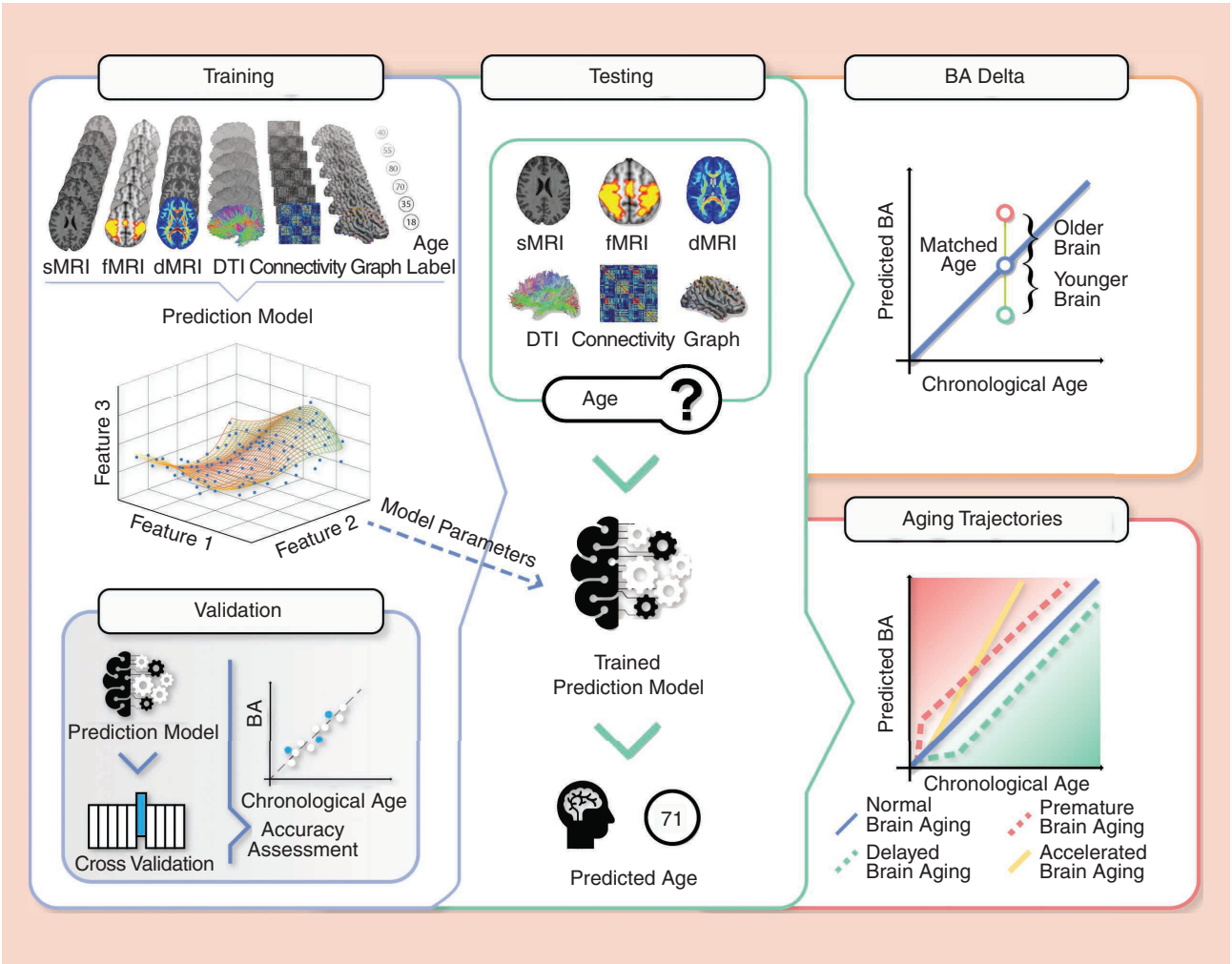

FIGURE 1. An overview of the BA prediction paradigm using supervised ML. Training: neuroimaging data as directly derived from multiple modalities or as secondary features from healthy individuals are labeled by participants' chronological age and given as input to a regression model. Validation: a k-fold cross-validation procedure is performed by dividing the training sample in k folds and using $k-1$ for training and hold-out for prediction. The procedure is then iterated through all the folds, and predicted ages are compared with chronological ones to assess the accuracy of the model. Testing: the model coefficients/weights resulting from training are applied to unseen participants' data to generate individual BA predictions. BA delta: the chronological age is then subtracted from the predicted BA to derive a single summary measure (the brain-PAD). Aging trajectories: differential trajectories of brain aging have been demonstrated, providing information about one's health status. Individuals can differ in their brain aging trajectories and deviate from what is considered "healthy aging." Indeed, a person may have genetic and developmental/environmental factors that lead to a higher rate of aging throughout life (premature aging; the red dashed line) or, conversely, delayed brain aging (the green dashed line). In addition, someone may experience specific events during that determine an accelerated trajectory of brain aging (the yellow line). sMRI: structural MRI; fMRI: functional MRI; DTI: diffusion tensor imaging.

[13]. These findings suggest that input brain features could have a greater impact than the model choice.

Concerning different IDPs, the most accurate studies in adults have reported an MAE of 4–5 years by using single-modal models based on T1w structural IDPs [15], [25], with slightly better results when using subcortical [17], [24] and voxel-based morphometry (VBM) features [13]. Moreover, in studies covering age ranges between early childhood and young adulthood, the predictions generally reached better accuracy, with MAE values of only 1–2 years [21], [26]. Similar results for healthy adults were found in [16], where the authors additionally highlighted the importance of evaluating the influence of several parameters on the BA estimation framework. In particular, the number of training samples was found to have the strongest impact on accuracy, while processing parameters, including methods for deriving structural features from T1w images and dimensionality reduction with PCA, had only a mild influence on MAE values.

In recent years, researchers have also attempted to explore the potentialities of additional brain IDPs to generate predictions of BA. Recent cross-sectional studies have investigated a wealth of different MRI data and corresponding IDPs on large samples of healthy adults either by simply combining all features [3] or using structural MRI results as a benchmark. These latter studies revealed that single-modal predictions based on dMRI tended to achieve performance comparable to those using conventional T1w data, while models based on fMRI measures were able to explain only a limited amount of age variance and reached the lowest prediction accuracy [4], [20]. Moreover, only moderate performance was shown by CBF-based models in the study by Rokicki et al. [24], which is the only one comparing the prediction accuracy of ASL features with that resulting from well-known structural measures.

Besides single modalities, all these cross-sectional studies investigated whether merging information from different MRI data in a multimodal model could help achieve more accurate BA predictions. The models integrating different modalities and feature sets led to the best fit in all these studies [4], [20], [21], [24]. Similar findings resulted from using multimodal stacked models [17], [23], which were able to predict BA better than using MRI alone and helped gain almost one year of error when compared to purely anatomy-based prediction when stacking either MEG or fMRI on the basic model.

While the importance of merging multiple features to achieve accurate BA estimates is undeniable, Smith et al. [10] recently remarked how the combination of several factors into a single metric is useful but might come at the cost of losing important information for distinguishing different biological factors contributing to changes seen in the brain while aging. Therefore, besides feeding all the available IDPs into a single model, which achieved the best accuracy (MAE = 2.9 years), the authors analyzed different aspects of brain aging represented by 62 modes, each reflecting a combination of IDPs, and derived a series of modes' BA estimates per subject. Their findings demonstrated the added value given by considering multiple modes separately and highlighted the importance of going beyond a single summary measure to reveal more biologically meaningful brainprints of aging.

DL

The quest for more accurate age estimates and increasing availability of big neuroimaging data have pushed researchers to go beyond conventional ML approaches and leverage the potentialities of DL, even in the BA framework. In the past year, a few studies on the topic have been published, as we briefly summarize in the "Key Results From BA Studies Using DL Methods" section, showing promising results when using 3D T1w images as inputs and achieving SOA MAE values. However, challenges exist for further improving prediction accuracy, especially on small data sets with limited training samples, and for defining an optimal architecture. Indeed, whether more complex and deep models perform better than simpler ones in the BA task and how they behave with different types of data are open issues to be elucidated.

Six of the seven relevant papers included here relied on a CNN architecture to estimate BA using 3D T1w images (for more details about the CNN, please refer to the supplementary material (available in https://doi.org/10.1109/MSP.2021.3126573) with this article.). While numerous variants of CNNs are present in the current literature, the solutions explored in the BA context so far are mostly based on Visual Geometry Group Network (VGGnet) and residual NN (ResNet) architectures (Figure 2). As

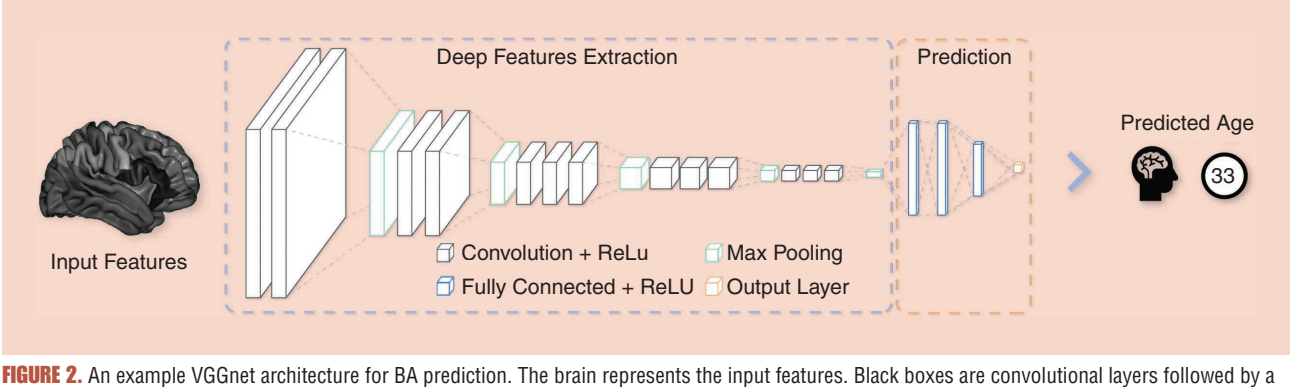

FIGURE 2. An example VGGnet architecture for BA prediction. The brain represents the input features. Black boxes are convolutional layers followed by a rectified linear unit (ReLU) activation layer; green boxes are max-pooling layers, while the blue box represents fully connected layers, followed by a ReLU activation layer. Finally, the orange box is the linear predictor. The output is the predicted age of the subject.

for ML-based BA pipelines, a data sample is split into training, validation, and test sets for model training and evaluation. During the training process, CNNs are usually optimized using stochastic gradient descent and its recent extension, adaptive moment estimation, while the MAE between true and predicted ages and the Kullback–Leibler divergence are used as a loss function to be minimized in the process.

In this last case, BA prediction is framed as a soft classification problem rather than classical regression, as recently proposed by [6]. In this setup, the label of the age is not considered a single number but a discretized Gaussian probability distribution centered at the true chronological age and with a distribution sigma set to be the size of one age bin, such as 1–2 years. The output of the model is also a probability distribution, and the Kullback–Leibler divergence can be used to measure the similarity between the two probabilities. The resulting output has a given number of X digits standing for X age bins, each covering a specific year range, and the final age prediction is given by the average of all the age bins weighted by the output probability. This soft classification approach should enforce a model to predict BA as accurately as possible [6].

To reduce model reliance on preprocessing steps, such as image realignment and registration, all the studies tend to apply only minimal preprocessing to the input data. In addition, different regularization and data augmentation strategies, including dropout, data rotation, translation, mirroring, scaling, and the addition of random noise, are usually applied during the training phase to avoid overfitting and improve generalization. Finally, the performance of a model is evaluated by MAE and r-values, as described in the "Statistical Methods/ML" section.

Key results from BA studies using DL methods

The past two years have witnessed an increase in the number of studies exploring CNN-based methods for age prediction from T1w images, enabling a major leap in the understanding of this problem. While in all the representative papers reviewed here the CNN architectures were mostly based on the two previously mentioned models, specific designs were proposed in each study to achieve accurate estimates and better tackle some of the main computational issues (e.g., the number of total parameters, complexity, and memory requirements). Inspired by the general VGGnet architecture, in [12], a 3D CNN was devised, taking as input whole-brain T1w volumes from a large aggregated data set covering the adult life span. The model performed well in both the hold-out test set from the same population and in another independent data set (MAE ~ 4 years), although it did not achieve the best prediction results compared to other studies using a DL framework. However, as the authors recognized, the main focus was not accuracy itself but the assessment of the contribution of age distribution (uniform/nonuniform) in the training set. The findings demonstrated that a uniformly distributed data set would lead to accurate estimations without appreciable bias toward a certain age group while maintaining good training efficiency, promoting the inclusion of such criteria in a wider variety of BA studies.

Conversely, a recently published study by Peng et al. [6] aimed at proposing a novel 3D CNN architecture to achieve the best performance possible, as confirmed by the authors' participation in the 2019 Predictive Analysis Challenge for BA forecasting, where they ranked first. They developed a lightweight DL architecture, the simple fully convolutional network (SFCN), based on a fully convolutional network and VGGnet characterized by a relatively low number of parameters and taking as input whole-brain 3D T1w images. The design they created enabled reducing the computational complexity and memory cost, and good accuracy values could be reached with the UKB data set, even with a low number of training subjects. They also compared their SFCN with more complex CNN architectures (e.g., ResNet18 and ResNet50), demonstrating that deeper models do not perform better than shallow ones in this prediction task and reaching the lowest MAE with their lightweight model. Finally, they demonstrated that, regardless of the training set size, the SFCN was able to outperform a well-tuned Elastic Net model.

That was in line with previous findings with a 3D CNN and T1w images by the authors of [13]. In the study, the authors started by independently training their model on four different structural features (whole-brain T1w images, Jacobian map, and GM/WM segmentations), achieving an MAE of 4–4.8 years, depending on the image type. In addition, they compared this approach to eight ML regression models separately trained on surface-based morphometry, VBM, and similarity matrix features, showing that the DL models predicted BA more accurately than the others. Similar promising MAE values using CNN models have been reported by other authors in healthy individuals from various databases [7], [8], [11]. Interestingly, Jonsson et al. [13] and Peng et al. [6] investigated whether adding together different sets of features extracted from T1w images, such as GM/WM maps and raw, nonlinearly normalized T1w data, could boost performance, as demonstrated for BA prediction with ML regression methods. Starting from single CNN models separately trained with different feature types, ensemble models were derived using a majority voting strategy to form the final prediction [13] and averaging the single predictions [6], and both led to performance gains through multimodal inputs when compared to using a single modality. This once again highlights the importance of combining multiple data types, given the somewhat independent information gathered from different features [6].

Finally, a few authors have started to assess the feasibility of transfer learning (TL) approaches to fine-tune a BA model initially trained on a data set acquired from different sites and under different imaging conditions, with promising results [13], [19]. Indeed, TL might be a plausible solution to improve the generalizability of MRI-based BA prediction models, especially when the large data requirements typical of ML/DL are not met and when dealing with more sophisticated measures, such as those derived from dMRI and fMRI data, which often experience greater intersite variability. An important example can be found in [19], where CNN models were applied to predict BA by using tract-based measures from dMRI data (tensor and more advanced features, such as GFA and NG). Once the models were defined, TL was applied to transfer the dMRI aging models from the source domain (Cam-CAN) to the target domain [NTUH (National Taiwan University Hospital), IXI-HH

(Hammersmith Hospital), IXI-Guys], representing four independent data sets with different acquisition schemes/parameters, which have a clear impact on microstructural dMRI maps and would preclude the generalization of a brain aging model. Their results indicated that the pretrained model built using the source domain could be transferred and fine-tuned to the three other data sets with satisfactory prediction performance and with good test–retest reliability.

Overall, the key points regarding BA prediction modeling can be summarized as follows:

1) BA models based on multiple statistical/ML approaches are still a common choice in most studies.
2) Literature findings using different data sets have shown that input features have a higher impact on BA estimates than does the adopted ML method and thus need to be carefully selected.
3) BA predictions based on DL approaches are emerging and bring added value in many respects, including the possibility to directly use raw neuroimaging data, avoiding feature engineering if a data set is sufficiently rich, the ability to cope with large amounts of data/features, the possibility to model nonlinear relationship between input and output variables, and higher accuracy in BA prediction compared to conventional ML approaches.
4) Multimodal approaches combining different feature types in either ML or DL frameworks generally achieve higher accuracy compared to single modality.

For additional information, refer to "A Glimpse Into Some Methodological Pitfalls in Brain Age Estimation."

XAI

XAI recently emerged as one of the hottest topics for understanding "the why and how" of the outcomes of ML/DL algorithms. However, this is still largely unexplored, especially in the brain aging field, though it could help to disentangle the contributions of different features shaping final estimates as well as provide other hints about aging mechanisms that cannot be captured with traditional approaches. Sixteen of the 24 selected papers employed XAI. Before tackling the issue, we try to elucidate one aspect that is still unclear in the literature, i.e., the difference between explainability and interpretability, which are used quite interchangeably while describing different concepts. Then, we give an overview of the most-used methods in the context of BA prediction, and we review related literature addressing the question: What is XAI, and how can it be used to assess and understand model outcomes?

Interpretability, causability, and explainability

The concepts of explainability and interpretability are hard to encode and usually considered interchangeable by ML researchers. Such an ambiguity was also put forth by the query outcomes of the literature review. In fact, the keyword *explainability* did not return any results, while *interpretability* returned the papers discussed in this section. A clear definition of such terms would be required to achieve agreement about their meaning in this context and derive criteria for their assessment, either subjective or objective. Far from pretending to solve this issue, which would require philosophical thinking, we shape the discussion around one possible signification.

Following [29], interpretability is connected with the human intuition behind the outputs of a model, claiming that the more interpretable the model is, the easier it is to devise cause-and-effect relationships within the system input and output. This definition is strongly related to the the concept of causability, which is quite relevant to the medical area and presented in [30]. *Causability* is defined as "the extent to which an explanation of a statement to human experts achieves a specified level of causal understanding with effectiveness, efficiency, and satisfaction in a specified context of use." Instead, following [29], explainability would be associated with the decoding of the internal logic and mechanisms of an ML system. In particular, the authors of [30] define *explainability* as highlighting the decision-relevant parts of the representations of an algorithm and active parts in the algorithmic model that contribute to model accuracy on a training set or a specific prediction for a particular observation. It is hence not necessarily related to human understanding.

Therefore, regarding ML, interpretability does not axiomatically entail explainability and vice versa, following [29]. Figure 3(b) tries to express the difference between these two concepts. Starting from the training data, two directions can be followed to obtain model explanations: 1) using a directly explainable model, such as a decision tree or linear regression model, for which the underlying logic is easy to follow and understand and the explanation can be straightforwardly derived from the model coefficients, and 2) applying a black-box model (e.g., deep models, such as CNNs), followed by a post hoc interpretability model to derive explanations, not necessarily requiring an understanding of the underlying model mechanism.

For a system to be interpreted, explanations, namely, the outcomes of interpretability method applications, must be provided, and the properties making an explanation effective for humans need to be defined. Holzinger et al. [30] state that data, objects, and graphical representations $\leq \mathbb{R}^3$, such as images and text, are directly understandable and hence interpretable by humans. Feature probing methods provide explanations to enable model interpretation. Following [31], three feature properties are relevant: 1) feature stability assessed through approaches that measure how stable each feature contribution is across multiple models trained on held-out data sets by using resampling methods and cross-validation, 2) a ranking of feature importance obtained by assessing the impact of a feature on the prediction output, and 3) feature visualization that encompasses strategies providing a visual rendering of importance, such as saliency maps. While the second and third properties aim at making model outcomes human understandable, the first one can be considered a way to assess the robustness (generalizability) of a solution. In this respect, bootstrapping is usually employed in the training/validation phase.

In [32], the author discussed the properties that might render models interpretable, highlighting that human decisions might admit post hoc interpretability despite the black-box nature of human brains. One advantage of this reading of interpretability

is that opaque models can be interpreted after the fact, and it subtends a clear distinction with respect to explainability, which instead entails the clear understanding of a model's internal rules and functioning. In our work, we build on such a claim and assume that interpretability points to causability, while explainability means decoding a system's internal rules; the two do not reciprocally entail and fall under the XAI umbrella. In addition, only data types strictly related to BA prediction, including images, graphs, and tabular data, are considered. The reader can refer to [29] for a broader discussion of interpretability methods for different data types.

In what follows, we review a few additional attributes of interpretability models that we consider relevant in this context. Interpretability models can be model agnostic or model specific. While the former tries to give some insights about the function underlying a model, regardless of the model structure, the latter can be applied only to a specific prediction model and architecture. Moreover, interpretability models can be local or global depending on whether the explanation concerns an individual prediction or small sections of a whole model or system, respectively. Finally, another pertinent categorization proposed in [30] distinguishes between post hoc and ante hoc models. In our taxonomy, the first group lies in the interpretation models, while the second involves explainable models that also have the interpretability property, as they embed explainability directly into their structure. In what follows, these categories, represented in Figure 3, are briefly discussed along with our main findings in the BA context.

A Glimpse Into Some Methodological Pitfalls in Brain Age Estimation

The following two important aspects have to be kept in mind while designing an individual brain age (BA) model:

1) *Age bias correction*: Several studies are reporting age-dependent bias in predictions, which contributes to interpretation uncertainty. What has been observed is that BA tends to be biased toward the mean age of the training cohort, meaning that BA is overestimated in younger subjects and vice versa in older ones. This is due to regression dilution and other factors, such as model regularization and non-Gaussian age distribution [3], [6], leading to a significant dependence of the brain-PAD on chronological age.

   To overcome this issue, statistical bias correction is generally applied to a predicted BA or brain-PAD estimates through post hoc linear methods. The most common practice is to calculate the regression line between the chronological age (the predictor) and brain-PAD (or the predicted BA, outcome) in the training set and then use the slope and intercept to derive the corrected predictions in the test set. As a result, the brain-PAD and chronological age are not correlated anymore. These procedures have been successfully employed in different studies, relying on both machine learning [4], [10], [20] and deep learning (DL) models [6]. Other approaches could be devised to embed the correction directly in the regression models, as, for example, in [14], where the authors proposed a modification of three popular regression models (ordinary least-squares, Ridge, and kernel Ridge regression), introducing an additional hyperparameter to control the maximum possible correlation between the brain-PAD and true age. Therefore, attention should be paid to this point when comparing results across studies [S1].
2) *Confound modeling*: Confounds might introduce spurious associations between independent variables and strongly bias the resulting estimates. The criteria defining the role of confounds are far from trivial and heavily context dependent. In the BA framework, for instance, age plays the role of the variable of interest but also of a possible confound. These variables can be assigned to common categories depending on their nature, such as subject-specific features [e.g., age, gender, education, intracranial volume, and Apolipoprotein E (APOE)], scanner/acquisition/processing parameters (e.g., the center, coil, and head motion) and nonlinear/interaction terms [S2]. The most common approach relies on basic variables, such as gender and sites, while a richer set of covariates was found in studies relying on the UK Biobank [3], [10]. Once the confounds are defined, the common practice is to regress them out from the data as a preprocessing step (deconfounding) or add them as regressors in all the analyses.

   In the specific case of DL and 3D convolutional neural networks (CNNs), besides the deconfounding strategy, two main approaches were observed: 1) testing with a linear regression model the main effects of the covariates on the cross-validated brain-PAD estimates [26] and 2) adding covariates as inputs to the final CNN layer [13]. These play the role of additional constraints during training to limit the solution space of the network, forcing the net to more accurately capture the relevant factors and their interactions. All these aspects deserve further investigation, as they would have a great impact on the statistical power of the analysis as well as the outcomes of the association analyses.

References

[S1] A.-M. G. de Lange and J. H. Cole, "Commentary: Correction procedures in brain-age prediction," *Neuroimage, Clin.*, vol. 26, p. 102,229, 2020, doi: 10.1016/j.nicl.2020.102229.

[S2] F. Alfaro-Almagro, P. McCarthy, S. Afyouni, J. L. Andersson, M. Bastiani, K. L. Miller, T. E. Nichols, and S. M. Smith, "Confound modelling in UK biobank brain imaging," *Neuroimage*, vol. 224, p. 117,002, Jan. 2021, doi: 10.1016/j.neuroimage.2020.117002.

Ante-hoc models

This class includes explanability models, such as linear, decision tree, and rule-based models, and more complex models that are equally transparent [29]. Six of the 16 papers applying XAI exploited ante hoc models to give explanations for their outcomes. In BA prediction, as presented in the "BA Prediction Modeling: From Statistical Methods to DL" section, linear regression models, such as OLS, Ridge, and LASSO, are widely used. They hold the interpretability property, in that they can be directly interpreted in terms of their β coefficients both locally and globally. Assuming that the data have been standardized and the model contains no intercept, large components of β can be interpreted as features that are relevant to the regression task. It is noteworthy that interpretability is affected by the presence of collinear independent variables since, in this case, large weights could be assigned to features that are not related to the target variable. Structure coefficients and activation patterns have been proposed as alternative metrics [14]. Constraints, such as nonnegativity and regularization, that promote sparseness could be applied to restrict the set of solutions, potentially simplifying a model and improving interpretability.

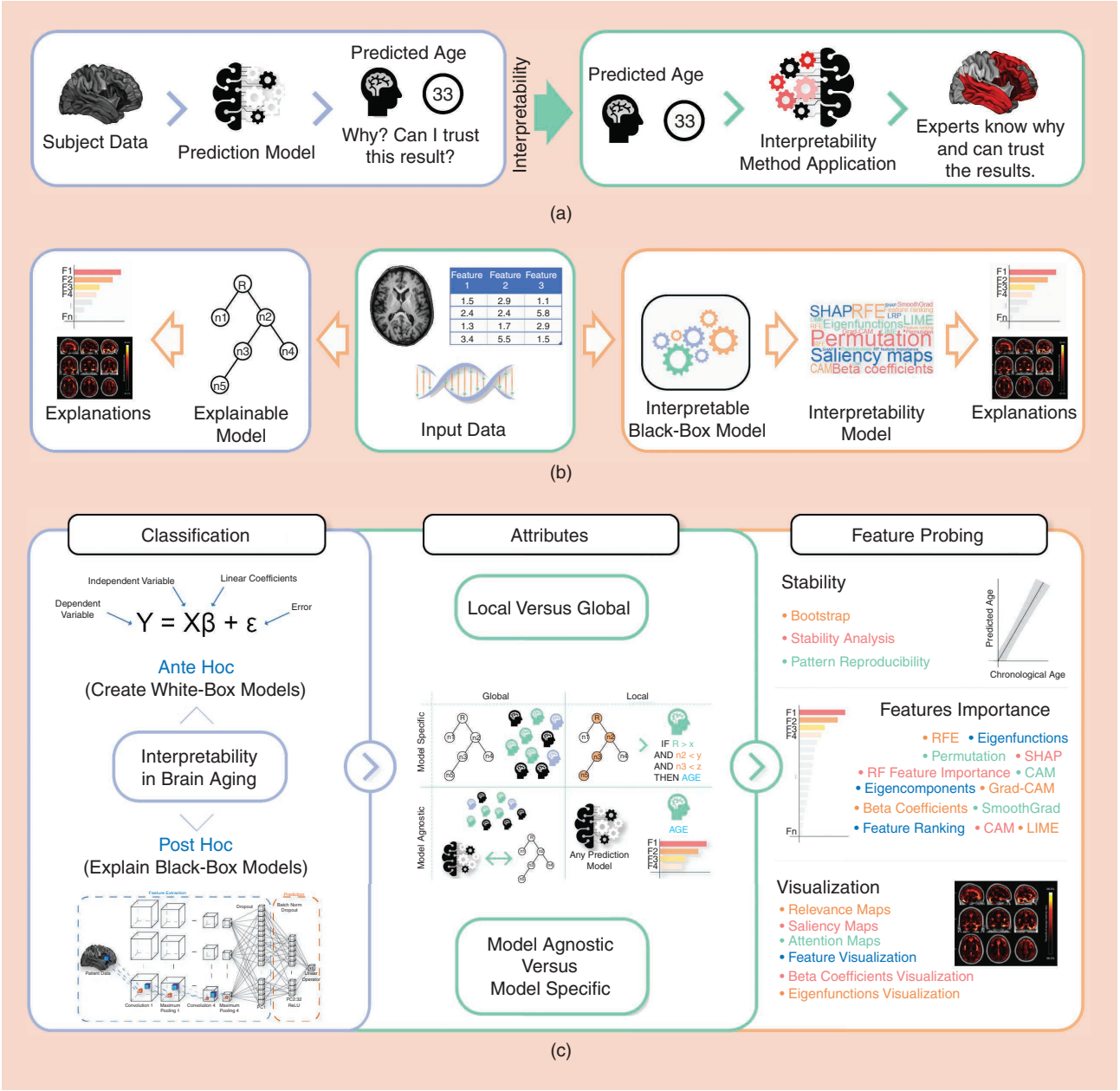

FIGURE 3. An overview of interpretability methods. (a) Black-box models can obtain optimal predictions, but they do not facilitate complete understanding. The application of interpretability methods enables retrieving and interpreting the most important features. (b) Schematic representation of the difference between explainable and interpretable models. (c) Interpretability methods classified as ante hoc and post hoc. Local, global, model-agnostic, and model-specific attributes are exemplified as well as feature-probing properties. SHAP: Shapley additive explanation; Grad-CAM: gradient-weighted class activation mapping; LIME: local interpretable model-agnostic explanations; LRP: layer-wise relevance propagation; RFE: recursive feature elimination.

Linear regression models are sometimes preceded by linear latent variable models, such as CCA, PCA, and ICA, or their generalization to perform an initial feature selection or find the "modes" that embed information derived from either single or multimodal data in a smaller feature space. Latent variable models yield loading vectors for every component, quantifying the contribution of each feature to each component. Thus, when applied to BA prediction, they facilitate understanding which feature mostly exhibited age-related changes. Both the β coefficients and loadings can be directly visualized in a feature space that, specifically for BA prediction, can be a brain map.

Post-hoc models

Post hoc models represent the widest class of interpretability methods and were applied by 10 of the selected papers implementing XAI. Concerning ML models, feature ranking based on feature permutation is often used as an interpretability technique. The permutation feature importance is a model inspection technique that can be used for any fitted estimator when data are tabular. The feature importance is defined as the decrease in a model score when a single feature value is randomly shuffled. The drop in the model score is indicative of how much the model depends on that feature. This technique benefits from being model agnostic and can be calculated many times with different permutations of the features.

Widely used interpretability methods for black-box models are perturbation-based methods, such as the Shapley additive explanation (SHAP), which aims at building surrogate models for black-box ones to provide them interpretability. The SHAP leverages the idea of Shapley values for model feature influence scoring, considering all possible predictions for an instance, using all possible combinations of inputs. Because of this exhaustive approach, the SHAP can guarantee consistency and local accuracy.

Moving to DL, saliency maps play a central role. The basic approach is to rely on gradients, each quantifying to what extent a change in each input dimension would modify the predictions in a small neighborhood around the input. The output is an image-specific class saliency map corresponding to the gradient of an output neuron with respect to the given class. Starting from this common point, multiple methods were developed to generate saliency maps from DL models. The most popular one in BA prediction is SmoothGrad. It first distorts the input image by adding normally distributed noise and then the extracts the partial derivative in each voxel with respect to the trained model output. This procedure is repeated multiple times to obtain an average saliency map derived from the multiple gradient maps.

A similar approach is class activation mapping (CAM), a method developed for CNNs and indicating the discriminative regions of an image used by a CNN to obtain an output. In detail, a feature vector is created by computing and concatenating the averages of the activations of convolutional feature maps that are located just before the final output layer. A weighted sum of this vector is fed to the final layer. The importance of the image regions can be assessed by projecting back the weights of the output layer on the convolutional feature maps. Gradient-weighted CAM (Grad-CAM) was built on top of CAM and can be applied to all CNN models, achieving improved object localization in the resulting heatmaps for any architecture.

Finally, ablation and occlusion procedures are important interpretability methods for DL. They iteratively cover a part of the input image with a black patch, and the network output is recalculated to assess changes in the prediction probability, assuming that the covered region was relevant for the forecast. Concerning the quality of saliency maps, a word of caution is needed. When applying such methods, relying solely on visual assessment to judge the results can be misleading. Tests based on statistical randomization, comparing a natural experiment with an artificially randomized experiment, can be applied to select the most suitable and reliable saliency method.

BA interpretable models

In BA prediction, a few interpretability models are starting to be employed to obtain a deeper understanding of the main factors leading to a predicted BA. They include the following:

1) *Ante hoc models*: Ante hoc explainable methods are widespread in this field. Five main contributions could be found in this category:
   - *Linear models*: Treder et al. [14] exploited regression coefficients to explain model predictions, demonstrating that these were able to capture the entire BA prediction pipeline (see the "BA Prediction Modeling: From Statistical Methods to ML" section).
   - *Latent variable models*: The common derivation poses latent variable models as feature reduction methods, as discussed in the "Neuroimaging-Based Phenotypes for BA Prediction" section. Latent space components enhance the interpretability of a model. In [22], PCA was employed as a preliminary step for a linear model relying on fMRI features. The latent variables corresponded to FC networks describing connectivity patterns. A similar approach was presented by Smith et al. [10], which used PCA and ICA to extract 62 modes of subjective variability, acting as prints for aging brains. Each mode represented different aspects of brain aging, showing distinct patterns of functional/structural brain changes as well as selective associations with genetics, lifestyle, cognition, physical measures, and disease.

     In the scalar-on-image procedure proposed by Palma et al. [18], the obtained eigenfunctions encoded the main differences among healthy, MCI, and AD subjects. The authors observed, for instance, that the first eigenfunction facilitated distinguishing the lateral ventricles from the rest of the brain, concluding that the scores for this eigenfunction could be correlated with diagnosis and chronological age. Finally, the authors of [17] directly employed CCA to visualize features significantly contributing to BA prediction.
   - *Stability assessment*: Cole [4] used the bootstrapping technique on LASSO coefficients to highlight the most informative variables among the multimodal ones included in his work, concluding that the multimodality model outperformed single-modality models. This is an indication that

multimodality improves performance, though much of the age-related variation can be captured by a small subset of features, including both T1w and dMRI endophenotypes.

2) *Post hoc models*: Even if post hoc models are still quite underexplored, we include 10 contributions to give a complete overview of the techniques applied so far:
   - *Permutation-based feature importance*: Starting from methods using a single modality, the authors of [11] used permutation feature importance to analyze the relevance of different WM/GM regions to predict BA for a healthy cohort. They highlighted regions playing a central role in cognition and memory (the hippocampus), emotional regulation/salience (the amygdala) and physiological homeostasis (the insula). Conversely, De Lange et al [20] and Rokicki et al [24] both developed a multimodal BA prediction model. While the authors of [20] exploited permutation-based feature importance to perform an initial feature reduction, the authors of [24] used this technique to actually interpret outcomes, revealing that the model integrating all modalities was mostly driven by the cortical thickness, T1w/T2w ratio, and subcortical volumes.

     A similar approach was followed in [21], where feature importance was calculated across the reduction of the R^2 of the regression model. Finally, Engemann et al. [23], in their stacked model, studied variable importance unraveling in the presence of an additive component between MEG and fMRI phenotypes. Their feature importance analysis suggested that what renders MEG nonredundant with regard to fMRI are regional differences in the balance of fast brain rhythms, as it is reasonable to expect.
   - *Perturbation-based feature importance*: In [33], the authors assessed the specificity of spatial brain-PAD patterns by training prediction models, each time using only a subset of features derived from occipital, frontal, temporal, parietal, cingulate, insula, and cerebellum regions, respectively. They highlighted differential spatial patterns across the 11 different clinical groups that were analyzed. This analysis suggested that relative aging across regions showed opposite patterns in neurodevelopmental (schizophrenia) versus neurodegenerative [multiple sclerosis (MS) and dementia] disorders.

     Only one work exploited the SHAP method to assess feature relevance in an adolescence cohort [26]. The authors used the SHAP on multiple ML methods, revealing that anatomical changes in a common set of regions drive model predictions of age, regardless of the model type. The regions found by following the most important features reflected developmental growth patterns of the cortex in childhood and adolescence.
   - *Saliency maps*: Concerning gradient-based methods, the authors of [7] exploited SmoothGrad to produce explanation maps for their CNN model. They analyzed 15 different databases, including healthy and pathological cohorts, and were able to create aggregated population-based explanation maps. The similarity between each pair of group explanation maps was assessed, and clustering was applied to highlight brain regions that contributed the most to age prediction. Such regions showed the highest correlation to the brain-PAD, indicating the specificity of the derived maps to their model.

     Finally, both Wang et al. [8] and Feng et al. [12] retrieved Grad-CAM attention maps from their T1w-based CNN to show the relative importance of different regions for BA prediction. In detail, the work in [8] found that while the network looked at the entire GM, the attention pattern was quite complex, suggesting that the brain-PAD is more related to specific features than to global measures of GM volume when predicting BA. Feng et al. [12] moved a step forward. Besides post hoc, Grad-CAM-derived saliency maps, they applied ablation analysis methods focusing on part of the input data. They highlighted patterns of neuroanatomical contributions of normal aging, providing evidence for the prominence of frontal regions in all age epochs in the adult life span.

In conclusion, we can state that interpretability, causability, and explainability are not synonymous and provide different views of a model. The key aspects of XAI adoption in BA estimation can be summarized as follows:

1) Despite the fact that explainable models, such as linear regression and regularized regression, are widely used in BA estimation, their respective explanations are still not pervasively derived.
2) Among the most recent papers, visualization of the model coefficients in the latent space is being increasingly exploited.
3) Permutation- and perturbation-based feature importance have a central role in extracting explanations, especially those employed for ML model interpretation.
4) Saliency maps, even if still highly underexplored, are starting to be used in accordance with the spread of DL models and mostly rely on gradient-based methods.

Association studies with BA findings

In this section, we present the main papers on association studies between BA findings and several imaging/nonimaging variables (e.g., clinical, environmental, and genetic factors), with the aim of providing evidence to support the relevance of this novel aging measure. The biological pertinence of brain-PAD measures is indeed still under debate. Considering that BA estimation relies on using the residual prediction error (i.e., residuals from statistical analyses) as a summary measure, some authors argue that grounding interpretations on this could be flawed, as more accurate models would ideally reduce this value [25]. A crucial element of determining their relevance lies in external validation, in particular, associating this index with other characteristics measured in the same individuals and holding clear aging significance. This could be related, for instance, to cognitive, physical, lifestyle, and genetic domains.

As an initial attempt, a few authors have studied associations between brain-PAD values and multiple imaging features, such as cortical thickness [12], and volumetric, tract- and FC-based measures [3]. The former, in particular, provided evidence that the brain-PAD was significantly associated with the thickness of

several cortical regions, including areas in the frontal/temporal lobe and the precuneus. These share a prominent role in normal aging, which supports the soundness and relevance of the brain-PAD. However, the increasing availability of large repositories, including information other than neuroimaging data, brought great benefits to the field, facilitating studying the association with more detailed and variegated variables, as is the case for the UKB repository. This data set includes extensive nonimaging data, such as genetics, lifestyle factors (e.g., exercise, smoking status, and alcohol intake), cognitive/biomedical measures (e.g., fluid intelligence scores, trail-making tests, systolic/diastolic blood pressure, weight, and body mass indexes) along with health outcome information from the health-care system [10].

The authors of [3] and [4] were among the first to explore this in-depth resource and assess the relationship between their estimated brain-PAD values and more than 5,000 nonimaging variables, relying on simple correlation analysis and linear regression models accounting for different confounds and multiple comparisons. Interestingly, the authors demonstrated that an older-appearing brain was associated with higher systolic/diastolic blood pressure, a history of stroke, a diagnosis of diabetes, smoking, and alcohol intake. Of note, the same results for the significant positive association between some of these cardiovascular risk factors and the brain-PAD were also confirmed in another study on a different data set [20], providing evidence of the generalizability of these findings. Several measures from the cognitive testing were also significantly associated with accelerated brain aging, in the direction one might expect in all cases. These results for cognitive performance are in line with a previous study [13] evidencing how the brain-PAD was related to worse performance on three important cognitive tests (digit substitution, trail-making, and reaction time tests), thus proving that this summary metric is highly related to cognitive impairment.

Finally, Engemann et al. [23] similarly found a correlation between the brain-PAD and cognitive fitness of individuals in the Cam-CAN data set, with a higher brain-PAD associated with lower scores on memory and speed-thinking tasks. The brain-PAD association with a selection of traits from the UKB was also recently investigated in [11], though relying on a Mendelian randomization analysis to study the potentially causal nature of some associations. The results demonstrated a causal relationship between increased diastolic blood pressure and older-appearing brains, suggesting that preventing hypertension would have a positive impact on brain aging. In addition, relationships between the brain-PAD and several clinical scores have been evaluated in different groups of patients, such as the Expanded Disability Status Scale (EDSS) in MS [9] and Mini-Mental State Examination (MMSE) in MCI and AD [5], [15], [18] (please refer to the "Neurodegenerative diseases and BA" section for more details). The studies' significant findings further underline the importance of the brain-PAD as a candidate biomarker of aging and prognostic tool to track disease progression.

Besides this converging evidence, in three recent studies on UKB data, the authors took a step ahead and proposed to use the brain-PAD as an informative phenotype for genetic association studies. Indeed, as the physiological underpinnings of this measure are likely to be diverse, genetic analysis offers the opportunity to explore factors that influence phenotypic variation, enabling a deeper understanding of the brain aging process and its longitudinal changes [33]. In [13], a genome-wide association study (GWAS) of the brain-PAD yielded two different genetic sequence variants [also called *single nucleotide polymorphisms* (*SNPs*); in this study, they were rs1452628-T and rs2435204-G] that correlated with reduced sulcal width and WM surface area. Indeed, the GWAS facilitates correlation analyses between any variable, e.g., the brain-PAD, and a massive number of SNPs to highlight any significant existing association.

In addition, the authors of [13] investigated heritability through GWAS summary statistics analysis. They demonstrated that the brain-PAD is heritable, suggesting the capture of biologically relevant signals, in line with earlier findings from [33]. In this latter study, the authors performed further analyses to assess the overlap between the genetic substrate of the brain-PAD and common brain disorders, starting from the GWAS summary statistics for several diseases, such as AD, MS, and schizophrenia. Their results unveiled several significant independent loci that are specific gene positions on a chromosome, showing pleiotropy (i.e., the influence of one gene on many apparently uncorrelated phenotypic traits) between the brain-PAD and all the included disorders. Most of the identified loci were related to schizophrenia. This proved for the first time the presence of overlapping genes between the brain-PAD in controls and common brain disorders.

Similar analyses were also performed recently in [10], where the authors explored the rich information provided by 62 modes of brain aging (see the "Neuroimaging-Based Phenotypes for BA Prediction" section) and performed separate genetic analyses using different modes' delta estimates. Interestingly, while the single all-in-one brain-PAD estimations reported no significant results, the 62 GWASs of brain aging modes led to a total of 156 significant peak associations. Among these, a genetic association was found with rs429358, the SNP that determines whether the Apolipoprotein E (APOE) gene allele is E3 or not, and a major locus associated with several neurodegenerative diseases, including AD and MCI. These results confirm the importance of assessing the genetic influence of the brain-PAD but, at the same time, suggest that biological specificity might be diluted when generating a single metric, highlighting once again the added value of considering multiple and multimodal BA modes separately.

In summary, the key aspects concerning BA association studies are:

1) Validating the BA framework and related measures using other characteristics measured in the same individuals is essential to prove the biological and clinical meaning of the brain-PAD metric.
2) Positive significant associations have been demonstrated between the brain-PAD and some cardiovascular risk factors as well as lifestyle measures, meaning that older-appearing brains are associated, among others, with increased smoking/alcohol consumption and higher blood pressure.
3) Genetic analysis can offer the opportunity to dig into factors that influence variations in predicted BA estimates across individuals.

Neurodegenerative diseases and BA

In this section, we face the issue of the clinical relevance of BA estimates by presenting the most relevant contributions that have investigated the relationship between BA and the signatures of some common neurodegenerative disorders. Research for a marker of brain disorder onset is one of the possible translational applications for the BA model, which has been pursued in several studies, especially in the context of neurodegenerative diseases. Indeed, since these conditions are very likely to influence the rate of biological aging, BA represents an intuitive and easy way to assess the extent of this process and possibly disentangle age-related factors from disease-specific changes. As expected, in the large majority of these articles, BA estimates were derived for groups of patients with MCI and AD [5], [15], [16], [18], [24], [33].

In their pioneering work, Franke et al. [16] demonstrated a significantly higher brain-PAD in AD compared to controls from the ADNI database, indicating an older-appearing brain in AD subjects. Lowe et al. [5] forged a step ahead and grouped ADNI subjects as controls, stable/progressive MCI patients, and AD patients, further classified as carriers/noncarriers of APOE4. This, in particular, is an allelic form of the APOE gene that has been shown to be associated with AD onset. The estimated brain-PAD scores in healthy and stable MCI subjects proved to be significantly different from those in progressive MCI as well as AD subjects, suggesting brain modifications that accelerate the brain aging process in both categories. In the same study, the authors analyzed several longitudinal time points, showing that progressive MCI and AD patients had an aging rate that was significantly faster than that of the other two categories as well as for APOE4 carriers with respect to noncarriers. The authors also pointed out the existing association between the brain-PAD and cognitive scores by using the whole sample and the more accurate prediction of the conversion to AD using BA metrics rather than cognitive performance in both APOE4 carriers/noncarriers.

Similar findings were more recently reported by Palma et al. [18]. In addition, taking advantage of eigenfunction visualization, they showed how, depending on the disease phenotype, different regions contributed to the related brain-PAD (see the "BA Interpretable Models" section). Similarly, Rokicki et al. [24], using several BA models and permutation-based feature probing, showed how differences in BA estimates can characterize distinct pathophysiological aspects of neurodegenerative diseases. Indeed, they observed that the best accuracy in distinguishing controls from AD patients was obtained using brain-PAD values from global T1w images, while the values derived from CBF-based models were more discriminative for MCI and subjective cognitive impairment compared to controls.

Beheshti et al. [15] investigated BA prediction in relation to both AD and Parkinson's disease (PD). Their study evinced that brain-PAD values from GM-based models were greater in AD subjects than in PD subjects, while they were more similar in the two categories when derived from WM (although again higher in AD patients). Interestingly, the two models led to significantly different results for PD patients, hinting that multiple region-wise BA predictions might be useful for a better characterization of the aging process. The greater deviation from normality in AD patients suggests that the subjects have an older-appearing brain when compared to PD patients. The authors also carried out tests of the partial correlation between the brain-PAD and cognitive scores, resulting in significant associations only with the Montreal Cognitive Assessment test in PD subjects.

MS, often listed among the neurodegenerative disorders even though it is a multifactorial autoimmune disease, has been investigated in different studies. In [33], the authors included a great variety of brain disorders and reported increased brain-PAD values, with the strongest effects observed in schizophrenia, MS, MCI, and dementia. They also tested the same models using features derived from subgroups of ROIs to assess the specificity of spatial brain-PAD patterns across groups (more details can be found in the "BA Interpretable Models" section). Their findings demonstrated that regional estimates were largely comparable to those from full-brain analysis, with some notable differential spatial patterns across diseases. In particular, the brain-PAD calculated from cerebellar–subcortical features was higher in dementia and MS subjects. Moreover, association studies revealed that brain-PAD estimates derived from full-brain features in MS were correlated with the main clinical score used to assess the degree of disability (EDSS), while the ones derived from full-brain and cerebellar–subcortical features were correlated with cognitive functioning assessed with the MMSE in MCI and dementia, respectively.

Finally, Cole et al. [9] recently proposed an in-depth analysis of BA patterns in different MS phenotypes and showed a significant increase in the estimated brain-PAD in diagnosed MS patients compared to those affected by clinically isolated syndromes. Moreover, a longitudinal analysis revealed that increasingly greater brain-PAD estimates were positively correlated with worse disability evaluation scores (EDSS). Their findings also demonstrated the importance of this summary metric as a predictor of disability status and the presence of faster aging affecting relapsing–remitting MS patients compared to primary–progressive ones.

In summary, the main findings for BA in neurodegenerative diseases are:

1) A predicted BA and corresponding brain-PAD are biologically informative measures in different conditions, representing novel biomarkers to be further exploited.
2) Several disorders, including MCI, AD, and MS, demonstrated significantly higher brain-PADs (i.e., an older-appearing brain) compared to healthy controls.
3) Promising results are emerging by deriving multiple regional BA estimates using ROI-based features and assessing the specificity of spatial brain-PAD patterns across diseased groups.

Open issues and research directions

Brain aging is an exquisitely multidimensional topic that can be faced at different scales, from genes to behavior, and its assessment casts shadows across a multitude of domains, at both the clinical, including neurosciences, and societal levels. As highlighted in [2], the beauty of BA estimation is that it is an aging biomarker that summarizes complex information in a number

that is a very easily comprehensible measure, though with all its possible limitations discussed in [10]. This fits very nicely in the XAI framework, condensing the outcomes of the processing of diverse information in the instancing of a concept (aging) that subtends basic cognitive categories and thus conveys a clear message. Moreover, it could be integrated with other forms of "bodily ages" that are currently under investigation, for instance, based on biochemical and physiological measures, to provide a broader and more comprehensive assessment of the biological aging process [2]. As recognized by several authors, there are still relevant issues that deserve investigation. In the remainder of this section, we provide further research directions.

Novel methodologies within the BA framework

The emerging field of BA prediction is rapidly evolving, as confirmed by the increasing number of published papers we found on the topic in the past few years (see Supplementary Figure S1 in the supplementary material available on *Xplore* with this article). While different approaches have been explored with satisfactory results, further developments could be envisaged for each of the building blocks of this framework, starting from additional IDPs derived, for example, from novel microstructural indices [34], possibly representing the myelin content [35], and from graph-based network analyses of dMRI/fMRI/ASL data [36] as well as their joint modeling, which are still largely unexplored in the current literature. Moving to the prediction methods themselves, variational autoencoder embedding of multimodal data, as presented in [37], could improve the understanding of the association among input features through the analysis of the obtained latent space in the BA context and perform an intrinsic feature aggregation leading to an effective modality fusion. Moreover, by intrinsically dealing with missing data, they could enable exploiting different and more numerous databases.

Of note, building approaches to deal with multimodal data is an important aspect that should be further pursued in future studies, also in the context of DL, as done, for example, in [13], where the authors used a CNN model ensembling strategy and built a prediction model by combining the strengths of a collection of simpler base models. In particular, they ensembled CNNs trained on four T1w-based images by using both a majority voting scheme and a linear regression data blender to combine BA predictions and demonstrated that combining predictions in these ways can reduced overfitting and increase accuracy.

On the interpretability side, methods such as local interpretable model-agnostic explanations for feature ranking and Grad-CAM++ or layer-wise relevance propagation for feature visualization are still underexplored but have shown promising preliminary results in BA prediction [38] and MRI-based DL models [39]. Besides all these aspects, further investigations into specific elements along the BA cascade are still needed, such as the issues of age bias correction, appropriate confound modeling, external validation of the BA model, and defining the optimal age range for training. The increasing availability of large data sets, some of which cover the adult life span and different disorders, would help to further dig into these methodological issues and consolidate the relevance of BA estimates.

Digging deeper into genetics

As apparent from the survey in the "Neuroimaging-Based Phenotypes for BA Prediction" and "BA Prediction Modeling: From Statistical Methods to DL" sections, the exploitation of genetic determinants (GDs) is in its infancy. While a vast literature is available for association studies linking the brain-PAD to genetic variants, either as separate SNPs or blow down in polygenic risk scores, the inclusion of such information for modeling is mostly unexplored, though it could bring valuable insight into the relative weight of GDs versus non-GD features, such as IDPs, in BA prediction. XAI would be the enabling technology in this respect, as it provides tools to "open the box" of complex models, such as deep CNNs, and explain the processing outcomes as well as to rank features that contributed to results and quantify their influence.

The exploitation of GDs could also be fruitfully exploited in many other respects. As mentioned in [2], the telomere length holds the potential for being an aging biomarker. There is currently some evidence that the telomere length decreases with aging, though the link between these two variables is still to be elucidated. Assessing this in an indirect way, for instance, investigating the association between IDPs and the telomere length through Mendelian randomization, could shed some light on the topic. Along the same line, deoxyribonucleic acid methylation could play a role, a fact that was used to shape the so-called epigenetic clock. These studies could provide new features that help to prune the tree of the whole set of aging-related features and thus trace the best path from measures to the brain-PAD in the feature space constellation. Finally, blending information provided by XAI approaches in terms of regional features that are the most relevant in BA prediction with gene expression from the Allen Human Brain Atlas, consisting of the ribonucleic acid intensities of several genes related to central biological functions, might open new perspectives for a deeper understanding of specific genetic factors that affect brain aging.

Explainability, interpretability, and causability: Assessment methods

As mentioned in the "XAI" section, steps have to be taken by the XAI community to find common agreement about the concepts of explainability and interpretability in general and particularly in the BA field. Then criteria for their assessment could be derived. In this respect, the generalizability and robustness of explanations should be kept apart from the effectiveness of the interpretation primaries (e.g., saliency maps, feature rankings, and so on), and their relation with causability and causality should be elucidated. As pointed out in [32], the desiderata, or objectives, of interpretability typically regard contexts where standard ML problem formulations are imperfectly matched to the complex tasks they are meant to solve. ML models are optimized to make associations, while researchers often use them in the hope of inferring properties of observed phenomena, especially in the medical field. However, associations learned by supervised learning algorithms are not guaranteed to reflect causal relationships. Disentangling causation from correlation in association studies is one of the great

challenges to be faced for empowering the traslational potential of the proposed methods.

Enabling technologies for secure and distributed data processing

Complex models dealing with multidimensional heterogeneous data require large data samples. As shown in the literature, the reproducibility of results is strongly affected by the sample size and balance across potential sources of bias (e.g., acquisition modalities and protocols) as well as data protection issues. Federated learning equipped with encryption protocols, such as homomorphic encryption, could provide a means for sharing data and models, hitting the target of big data collection while getting rid of security issues.

Concluding remarks

The framework of the BA prediction paradigm has proved to be of great value in understanding elements that underlie an individual's biological age and therefore in characterizing different aging trajectories. At the same time, it provides a simple though effective means to capture novel insights into brain mechanisms, possibly enabling the timely identification of risk for future cognitive aging and age-related brain disorders. ML and DL approaches have shown great promise in this context, underscoring the importance of moving toward multimodal approaches. But at the same time, there is a need to further explore sets of specific BA estimates given by selective and regional ensembles of IDPs. The application of XAI is not only useful and necessary but it also represents a huge opportunity for BA prediction. On the one hand, XAI in linear and latent variable models aids direct detection and visualization in a human-friendly framework that illuminates the most important features, while on the other hand, it enables the application of complex and deep models, reducing their opacity by enhancing their trustworthiness.

Acknowledgments

This work was partially supported by Fondazione Cariverona (Bando Ricerca Scientifica di Eccellenza 2018, EDIPO project, grant 2018.0855.2019). Ahmed Salih is supported by the INVITE program cofinanced by the European Union within the Horizon 2020 Program and by Regione del Veneto. This article has supplementary downloadable material available at http://doi.org/10.1109/MSP.2021.3126573, provided by the authors. Ilaria Boscolo Galazzo is the corresponding author.

Authors

Ilaria Boscolo Galazzo (ilaria.boscologalazzo@univr.it) received her Ph.D. degree in neuroscience in 2014 from the University of Verona, Verona, 37134, Italy, where she is a temporary assistant professor in the Department of Computer Science. She is a coauthor of 29 publications in peer-reviewed journals and 60-plus contributions to conferences, and she is an associate editor of *IEEE Access*. Her research interests include imaging genetics, functional magnetic resonance imaging modeling, brain connectivity, and multimodal data integration. She is a Member of IEEE.

Federica Cruciani (federica.cruciani@univr.it) received her M.S. degree, cum laude, in engineering and computer science in 2019 from the University of Verona, Verona, 37134, Italy, where she is a Ph.D. student of computer science in the Neuroimaging Lab. Her research interests include diffusion magnetic resonance imaging modeling and analysis in clinical applications, machine learning and deep learning explainability applied to neuroimaging data, and heterogeneous data integration.

Lorenza Brusini (lorenza.brusini@univr.it) received her Ph.D. degree, Doctor Europaeus, in 2018 from the University of Verona, Verona, 37134, Italy, where she is a postdoctoral research fellow in the Department of Computer Science Neuroimaging Lab. Her research interests include neuroimaging, including diffusion magnetic resonance imaging modeling and data analysis in health and clinical applications as well as simulations, multimodal data integration, and brain–computer interfaces. She is a Member of IEEE.

Ahmed Salih (ahmedmahdeeabdo.salih@univr.it) received his M.Sc. degree in bioinformatics from Leicester University, U.K. In 2019, he was awarded a grant under the Horizon 2020 INVITE initiative to pursue his Ph.D. degree in computer science at the University of Verona, Verona, 37134, Italy, under the guidance of Prof. Gloria Menegaz. His research interests include brain aging exploiting machine learning techniques, explainability methods applied to neuroimaging data, and revealing the impacts of genetic and environmental factors on brain phenotypes.

Petia Radeva (petia.ivanova@ub.edu) received her Ph.D. degree in image processing, computer graphics, and artificial intelligence from the Universitat Automa de Barcelona in 1996. She is a full professor at the University of Barcelona, Barcelona, 08007, Spain, where she is the principal investigator for the Computer Vision and Machine Learning consolidated research group and a senior researcher at the Computer Vision Center. She has been a Research Executive Agency Future and Emerging Technologies-Open program vice chair since 2015 and an international mentor in the EIT Health Wild Cards program since 2017. She has been an International Association of Pattern Recognition fellow since 2015, one of the Catalan Institution for Research and Advanced Studies best scientists in Catalonia since 2015, and the recipient of several international awards. Her research interests are in machine particularly deep learning, computer vision, and applications to healthcare.

Silvia Francesca Storti (silviafrancesca.storti@univr.it) received her Ph.D. degree in neuroscience in 2012 from the University of Verona, Verona, 31734, Italy, where she is an assistant professor of bioengineering in the Department of Computer Science. She is a coauthor of 54 publications in international peer-reviewed journals on bioengineering and neurophysiological topics and 80-plus contributions to conferences. She is an associate editor of *BioMedical Engineering OnLine*. Her research interests include neuroengineering and focus on multimodal functional neuroimaging integration, brain connectivity inference, network analysis, and brain–computer interfaces. She is a Member of IEEE.

Gloria Menegaz (gloria.menegaz@univr.it) received her Ph.D. degree in applied sciences from the Swiss Federal

Institute of Technology Lausanne in 2000. She is a professor in the Department of Computer Science, University of Verona, Verona, 31734, Italy, where she leads the Neuroimaging Lab. She is an associate editor of *IEEE Signal Processing Letters* and *Frontiers in Digital Public Health* and a guest editor of *IEEE Signal Processing Magazine*. She is vice chair of the Research Executive Agency Future and Emerging Technologies-Open program and Marie Skłodowska-Curie Actions Individual Fellowships and an independent expert for international institutions, including the European Science Foundation. In 2004 she was awarded a Rita Levi Montalcini grant. Her research interests include neuroimaging, connectivity modeling, and imaging genetics, relying on machine/deep learning as enabling technology. She is a Senior Member of IEEE and a member of IEEE Women in Engineering and IEEE Women in Signal Processing.

References

[1] K. Franke and C. Gaser, "Ten years of brainage as a neuroimaging biomarker of brain aging: What insights have we gained?" *Frontiers Neurol.*, vol. 10, p. 789, Aug. 2019, doi: 10.3389/fneur.2019.00789.

[2] J. H. Cole, R. E. Marioni, S. E. Harris, and I. J. Deary, "Brain age and other bodily 'ages': Implications for neuropsychiatry," *Mol. Psychiatry*, vol. 24, no. 2, pp. 266–281, 2019, doi: 10.1038/s41380-018-0098-1.

[3] S. M. Smith, D. Vidaurre, F. Alfaro-Almagro, T. Nichols, and K. L. Miller, "Estimation of brain age delta from brain imaging," *Neuroimage*, vol. 200, pp. 528–539, Jun. 2019, doi: 10.1016/j.neuroimage.2019.06.017.

[4] J. H. Cole, "Multimodality neuroimaging brain-age in UK biobank: Relationship to biomedical, lifestyle, and cognitive factors," *Neurobiol. Aging*, vol. 92, pp. 34–42, Aug. 2020, doi: 10.1016/j.neurobiolaging.2020.03.014.

[5] L. C. Lowe, C. Gaser, and K. Franke, "The effect of the APOE genotype on individual BrainAGE in normal aging, mild cognitive impairment, and Alzheimer's disease," *PLoS One*, vol. 11, no. 7, p. e0157514, 2016, doi: 10.1371/journal.pone.0157514.

[6] H. Peng, W. Gong, C. F. Beckmann, A. Vedaldi, and S. M. Smith, "Accurate brain age prediction with lightweight deep neural networks," *Med. Image Anal.*, vol. 68, p. 101,871, Feb. 2021, doi: 10.1016/j.media.2020.101871.

[7] G. Levakov, G. Rosenthal, I. Shelef, T. R. Raviv, and G. Avidan, "From a deep learning model back to the brain–Identifying regional predictors and their relation to aging," *Human Brain Mapping*, vol. 41, no. 12, pp. 3235–3252, 2020, doi: 10.1002/hbm.25011.

[8] J. Wang *et al*., "Gray matter age prediction as a biomarker for risk of dementia," *Proc. Nat. Acad. Sci.*, vol. 116, no. 42, pp. 21,213–21,218, 2019, doi: 10.1073/pnas.1902376116.

[9] J. H. Cole *et al*., "Longitudinal assessment of multiple sclerosis with the brain-age paradigm," *Ann. Neurol.*, vol. 88, no. 1, pp. 93–105, Jul. 2020, doi: 10.1002/ana.25746.

[10] S. M. Smith, L. T. Elliott, F. Alfaro-Almagro, P. McCarthy, T. E. Nichols, G. Douaud, and K. L. Miller, "Brain aging comprises many modes of structural and functional change with distinct genetic and biophysical associations," *Elife*, vol. 9, p. e52677, Mar. 2020, doi: 10.7554/eLife.52677.

[11] A. Kolbeinsson, S. Filippi, Y. Panagakis, P. M. Matthews, P. Elliott, A. Dehghan, and I. Tzoulaki, "Accelerated MRI-predicted brain ageing and its associations with cardiometabolic and brain disorders," *Scientific Rep.*, vol. 10, no. 1, p. 19,940, 2020, doi: 10.1038/s41598-020-76518-z.

[12] X. Feng, Z. C. Lipton, J. Yang, S. A. Small, and F. A. Provenzano, "Estimating brain age based on a uniform healthy population with deep learning and structural magnetic resonance imaging," *Neurobiol. Aging*, vol. 91, pp. 15–25, Jul. 2020, doi: 10.1016/j.neurobiolaging.2020.02.009.

[13] B. A. Jónsson *et al*., "Brain age prediction using deep learning uncovers associated sequence variants," *Nature Commun.*, vol. 10, no. 1, p. 5409, 2019, doi: 10.1038/s41467-019-13163-9.

[14] M. S. Treder, J. P. Shock, D. J. Stein, S. DuPlessis, S. Seedat, and K. A. Tsvetanov, "Correlation constraints for regression models: Controlling bias in brain age prediction," *Frontiers Psychiatry*, vol. 12, p. 615754, Feb. 2021, doi: 10.3389/fpsyt.2021.615754.

[15] I. Beheshti, S. Mishra, D. Sone, P. Khanna, and H. Matsuda, "T1-weighted MRI-driven brain age estimation in Alzheimer's disease and Parkinson's disease," *Aging Disease*, vol. 11, no. 3, pp. 618–628, 2020, doi: 10.14336/AD.2019.0617.

[16] K. Franke, G. Ziegler, S. Klöppel, C. Gaser, and The Alzheimer's Disease Neuroimaging Initiative, "Estimating the age of healthy subjects from T1-weighted MRI scans using kernel methods: Exploring the influence of various parameters," *Neuroimage*, vol. 50, no. 3, pp. 883–892, 2010, doi: 10.1016/j.neuroimage.2010.01.005.

[17] A. Xifra-Porxas, A. Ghosh, G. D. Mitsis, and M.-H. Boudrias, "Estimating brain age from structural MRI and meg data: Insights from dimensionality reduction techniques," *NeuroImage*, vol. 231, p. 117,822, May 2021, doi: 10.1016/j.neuroimage.2021.117822.

[18] M. Palma, S. Tavakoli, J. Brettschneider, T. E. Nichols, and For The Alzheimer's Disease Neuroimaging Initiative, "Quantifying uncertainty in brain-predicted age using scalar-on-image quantile regression," *Neuroimage*, vol. 219, p. 116,938, Oct. 2020, doi: 10.1016/j.neuroimage.2020.116938.

[19] C.-L. Chen *et al*., "Generalization of diffusion magnetic resonance imaging–based brain age prediction model through transfer learning," *Neuroimage*, vol. 217, p. 116,831, Aug. 2020, doi: 10.1016/j.neuroimage.2020.116831.

[20] A.-M. G. de Lange *et al*., "Multimodal brain-age prediction and cardiovascular risk: The Whitehall II MRI sub-study," *Neuroimage*, vol. 222, p. 117,292, Nov. 2020, doi: 10.1016/j.neuroimage.2020.117292.

[21] X. Niu, F. Zhang, J. Kounios, and H. Liang, "Improved prediction of brain age using multimodal neuroimaging data," *Human Brain Mapping*, vol. 41, no. 6, pp. 1626–1643, Apr. 2020, doi: 10.1002/hbm.24899.

[22] R. P. Monti *et al*., "Interpretable brain age prediction using linear latent variable models of functional connectivity," *PLoS One*, vol. 15, no. 6, p. e0232296, 2020, doi: 10.1371/journal.pone.0232296.

[23] D. A. Engemann, O. Kozynets, D. Sabbagh, G. Lemaître, G. Varoquaux, F. Liem, and A. Gramfort, "Combining magnetoencephalography with magnetic resonance imaging enhances learning of surrogate-biomarkers," *Elife*, vol. 9, p. e54055, May 2020, doi: 10.7554/eLife.54055.

[24] J. Rokicki *et al*., "Multimodal imaging improves brain age prediction and reveals distinct abnormalities in patients with psychiatric and neurological disorders," *Human Brain Mapping*, vol. 42, no. 6, pp. 1714–1726, 2020, doi: 10.1002/hbm.25323.

[25] J. H. Cole and K. Franke, "Predicting age using neuroimaging: innovative brain ageing biomarkers," *Trends Neurosci.*, vol. 40, no. 12, pp. 681–690, 2017, doi: 10.1016/j.tins.2017.10.001.

[26] G. Ball, C. E. Kelly, R. Beare, and M. L. Seal, "Individual variation underlying brain age estimates in typical development," *Neuroimage*, vol. 235, p. 118,036, Jul. 2021, doi: 10.1016/j.neuroimage.2021.118036.

[27] P. Linardatos, V. Papastefanopoulos, and S. Kotsiantis, "Explainable ai: A review of machine learning interpretability methods," *Entropy*, vol. 23, no. 1, p. 18, 2021, doi: 10.3390/e23010018.

[28] A. Holzinger, G. Langs, H. Denk, K. Zatloukal, and H. Müller, "Causability and explainability of artificial intelligence in medicine," *Wiley Interdisciplinary Rev., Data Mining Knowl. Discovery*, vol. 9, no. 4, p. e1312, 2019, doi: 10.1002/widm.1312.

[29] L. Kohoutová, J. Heo, S. Cha, S. Lee, T. Moon, T. D. Wager, and C.-W. Woo, "Toward a unified framework for interpreting machine-learning models in neuroimaging," *Nature Protocols*, vol. 15, no. 4, pp. 1399–1435, 2020, doi: 10.1038/s41596-019-0289-5.

[30] Z. C. Lipton, "The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery," *Queue*, vol. 16, no. 3, pp. 31–57, 2018, doi: 10.1145/3236386.3241340.

[31] T. Kaufmann *et al*., "Common brain disorders are associated with heritable patterns of apparent aging of the brain," *Nature Neurosci.*, vol. 22, no. 10, pp. 1617–1623, 2019, doi: 10.1038/s41593-019-0471-7.

[32] M. Zucchelli, S. Deslauriers-Gauthier, and R. Deriche, "A computational framework for generating rotation invariant features and its application in diffusion MRI," *Med. Image Anal.*, vol. 60, p. 101,597, Feb. 2020, doi: 10.1016/j.media.2019.101597.

[33] L. Brusini, G. Menegaz, and M. Nilsson, "Monte Carlo simulations of water exchange through myelin wraps: Implications for diffusion MRI," *IEEE Trans. Med. Imag.*, vol. 38, no. 6, pp. 1438–1445, 2019, doi: 10.1109/TMI.2019.2894398.

[34] S. F. Storti, I. Boscolo Galazzo, S. Montemezzi, G. Menegaz, and F. B. Pizzini, "Dual-echo ASL contributes to decrypting the link between functional connectivity and cerebral blow flow," *Human Brain Mapping*, vol. 38, no. 12, pp. 5831–5844, 2017, doi: 10.1002/hbm.23804.

[35] D. Hu *et al*., "Disentangled-multimodal adversarial autoencoder: Application to infant age prediction with incomplete multimodal neuroimages," *IEEE Trans. Med. Imag.*, vol. 39, no. 12, pp. 4137–4149, 2020, doi: 10.1109/TMI.2020.3013825.

[36] A. Salih, I. Boscolo Galazzo, Z. Raisi-Estabragh, S. E. Petersen, P. Gkontra, K. Lekadir, G. Menegaz, and P. Radeva, "A new scheme for the assessment of the robustness of explainable methods applied to brain age estimation," in *Proc. IEEE Symp. CBMS*, 2021, pp. 492–497, doi: 10.1109/CBMS52027.2021.00098.

[37] F. Cruciani *et al*., "Interpretable deep learning as a means for decrypting disease signature in multiple sclerosis," *J. Neural Eng.*, vol. 18, no. 4, p. 0460a6, 2021, doi: 10.1088/1741-2552/ac0f4b.

SP