Medical Image Analysis 101 (2025) 103436

Contents lists available at ScienceDirect

Medical Image Analysis

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

A lightweight generative model for interpretable subject-level prediction

Chiara Mauri [a,b,*], Stefano Cerri [b], Oula Puonti [c], Mark Mühlau [d], Koen Van Leemput [b,e,f]

[a] Department of Health Technology, Technical University of Denmark, Denmark
[b] Athinoula A. Martinos Center for Biomedical Imaging, Massachusetts General Hospital, Harvard Medical School, USA
[c] Danish Research Centre for Magnetic Resonance, Center for Functional and Diagnostic Imaging and Research, Copenhagen University Hospital Hvidovre, Denmark
[d] Department of Neurology and TUM-Neuroimaging Center, School of Medicine, Technical University of Munich, Germany
[e] Department of Neuroscience and Biomedical Engineering, Aalto University, Finland
[f] Department of Computer Science, Aalto University, Finland

ARTICLE INFO

Dataset link: https://www.ukbiobank.ac.uk/

Keywords:
Image-based prediction
Brain age
Explainable AI
Generative models

ABSTRACT

Recent years have seen a growing interest in methods for predicting an unknown variable of interest, such as a subject's diagnosis, from medical images depicting its anatomical-functional effects. Methods based on discriminative modeling excel at making accurate predictions, but are challenged in their ability to explain their decisions in anatomically meaningful terms. In this paper, we propose a simple technique for single-subject prediction that is inherently interpretable. It augments the generative models used in classical human brain mapping techniques, in which the underlying cause–effect relations can be encoded, with a multivariate noise model that captures dominant spatial correlations. Experiments demonstrate that the resulting model can be efficiently inverted to make accurate subject-level predictions, while at the same time offering intuitive visual explanations of its inner workings. The method is easy to use: training is fast for typical training set sizes, and only a single hyperparameter needs to be set by the user. Our code is available at https://github.com/chiara-mauri/Interpretable-subject-level-prediction.

1. Introduction

Single-subject prediction methods aim to infer a subject's underlying clinical condition – such as their disease status – from its observed effect on the subject's anatomy or function as measured by medical imaging. The ability to perform this task accurately would have numerous potential applications in diagnosing disease, tracking progression, and evaluating treatment. It could also help clinicians to prospectively identify which patients are at highest risk of future disability accrual, leading to better counseling of patients and better overall clinical outcomes.

Many methods for automatic single-subject prediction have been proposed in the literature, using multivariate techniques that combine the weakly predictive power of many voxels simultaneously to obtain accurate predictions at the subject level (Arbabshirani et al., 2017; Cole et al., 2019). Recent years have seen a rapid growth in methods for predicting a subject's *age* from their brain scan, in particular, with the gap between the estimated and the real age being suggested as a potential biomarker of neurological disease (Cole et al., 2019; Kaufmann et al., 2019). Although high prediction accuracies can now be achieved, especially when methods are trained on the very large datasets that have recently become available (German National Cohort Consortium, 2014; Breteler et al., 2014; Schram et al., 2014; Miller et al., 2016; Alfaro-Almagro et al., 2018), comparatively little attention has been paid to *interpretability*, i.e., to the ability to explain the predictions to clinicians in terms that are biologically meaningful. Nevertheless, such interpretability is likely required before automated prediction methods can safely be adopted for widespread clinical use (Rudin, 2019).

A key difficulty in obtaining interpretability is that almost all subject-level prediction methods are currently based on *discriminative* learning, in which a direct mapping from an input image to a variable of interest is estimated from training examples. Especially with the deep neural networks that have become prominent in recent years, this results in "black box" models whose internal workings are hard to explain to humans. Although many of the post hoc explanation methods (Ras et al., 2022; Arrieta et al., 2020; Baehrens et al., 2010; Sundararajan et al., 2017; Springenberg et al., 2014; Selvaraju et al., 2017; Smilkov et al., 2017; Zeiler and Fergus, 2014; Bach et al., 2015) that have been developed for such complex models have raised specific criticism (Arun et al., 2021; Ghassemi et al., 2021; Adebayo et al., 2018; Rudin, 2019; Wilming et al., 2022; Sixt et al., 2020; Gu and Tresp, 2019), a more fundamental challenge is that interpretability is *intrinsically* hard for discriminative subject-level prediction methods

* Corresponding author at: Athinoula A. Martinos Center for Biomedical Imaging, Massachusetts General Hospital, Harvard Medical School, USA.
E-mail address: cmauri@mgh.harvard.edu (C. Mauri).

https://doi.org/10.1016/j.media.2024.103436
Received 9 June 2023; Received in revised form 6 December 2024; Accepted 9 December 2024
Available online 27 December 2024
1361-8415/

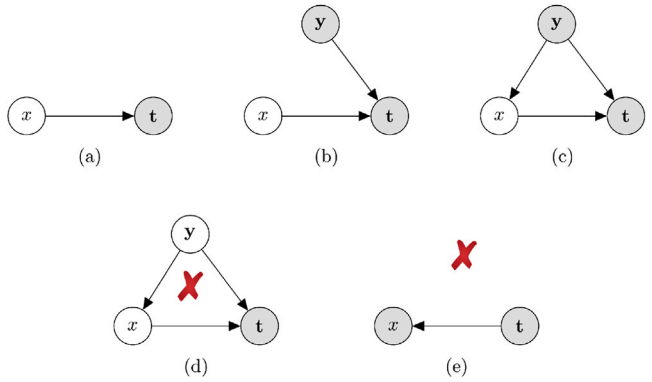


Fig. 1. The causal diagrams of the models considered in this paper are illustrated in the top row: (a) basic model encoding how the unknown variable of interest x generates an acquired brain scan **t**; (b) the model with additional known covariates **y** included; and (c) the model where **y** are known confounders. The arrows indicate causal relationships, and variables in empty vs. shaded circles are unknown vs. observed, respectively. The bottom row illustrates two cases that are not considered in this paper: (d) a model with confounders that are not observed; and (e) a decoding model where the direction of causality is reversed.

— even when very simple (e.g., linear) models are used. This is because discriminative methods optimize their prediction performance not only by amplifying the signal of interest in the data, but also by suppressing unrelated "distractor" patterns, so that their reason for looking at specific voxels cannot easily be deduced (Haufe et al., 2014; Wilming et al., 2022; Weichwald et al., 2015). Therefore, while e.g., linear discriminative methods are trivially transparent about *how* they compute their results (the weight they give to each image area can readily be inspected), they offer no explanation of *why* they are using specific image areas more than others (Ghassemi et al., 2021; Rudin, 2019; Haufe et al., 2014; Wilming et al., 2022; Lipton, 2018). Stated differently, their explanation refers to an understanding of how the model works, as opposed to an explanation of how the world works (Rudin, 2019).

Classical human brain mapping techniques, originally developed for analyzing functional images (Friston et al., 1991; Worsley et al., 1992; Friston et al., 1994; Worsley and Friston, 1995) but later adapted for structural imaging (Chung et al., 2001; Davatzikos et al., 2001; Wright et al., 1995; Ashburner and Friston, 2000; Fischl and Dale, 2000; Snook et al., 2007), are based on *generative* rather than on *discriminative* models: They encode a mapping from a variable of interest to the image domain, rather than vice versa. Their aim is to identify, on a population level, brain regions that are significantly correlated with specific variables of interest (such as disease status or an experimental condition) or their interactions. Especially when the variables of interest have a *causal* effect on brain anatomy – as is the case for e.g., age, gender or a particular brain disease – these methods are inherently interpretable: The spatial maps they provide indicate how each brain location would change, on average, if we had the ability to control the variable of interest at will.[1] However, because classical brain mapping techniques only consider each voxel-level measurement *independently* (so-called mass-univariate modeling), they are unsuitable for subject-level prediction, as each individual voxel by itself is typically only weakly predictive of the variable of interest.

In this paper, we propose a new method that aims to combine the predictive power of state-of-the-art multivariate models on the one hand, with the superior interpretability properties of classical brain mapping techniques on the other. This is accomplished by generalizing the independent, voxel-wise noise model in those classical techniques with one that also takes into account correlations *between* voxels. In particular, we use a linear-Gaussian latent variable model that allows us to simultaneously capture the dominant spatial correlations in the noise, control the number of free parameters that need to be learned from training data, and efficiently "invert" the model to make accurate subject-level predictions. Because it inherits the cause–effect relation modeling from classical brain mapping techniques, the approach is inherently interpretable: It provides both population-level spatial maps of the average effect of variables of interest on brain shape, as well as the ability to apply these effects at the level of individual subjects (so-called counterfactuals). The method is flexible and easy to use: It has only a single hyperparameter that needs to be tuned by the user, training is typically fast even without hardware acceleration, and incorporating additional covariates (and their possible interactions with the variable of interest) is straightforward.

An early version of this work appeared in Mauri et al. (2022). Here we significantly expand on the basic algorithm described in that paper, conducting an in-depth analysis of both the interpretability and the prediction performance of the method, detailing a fast practical implementation, and introducing extensions where dependencies on the variables of interest are nonlinear.

2. Context and related work

2.1. Problem statement

Fig. 1(a) illustrates the problem we address in this paper in its most basic form. An unknown quantity x in a patient is causing anatomical changes in their brain scan **t**. By modeling the causal relationship between x and **t**, we aim both to estimate the unknown value of x from **t** and to provide intuitive explanations of the estimation procedure — for instance by synthesizing images at different values of x and letting the user visually compare those with **t** (see Fig. 14 for examples). Models of the type shown in Fig. 1(a) are known as *generative* models in the machine learning community, or as *encoding* models in the neuroimaging literature (Friston et al., 2008; Weichwald et al., 2015).

For the purpose of estimating the parameters of the model, we need access to a set of training images for which the underlying value of x is known. It is worth mentioning that, unlike in age-prediction experiments where there is an abundance of available training data, many scenarios encountered in practice will provide only a few thousand training subjects at best: Imaging datasets collected for studying specific diseases typically have at most 1000–3000 subjects (Jack et al., 2008; Di Martino et al., 2014; Ellis et al., 2009; Satterthwaite et al., 2014), whereas even the largest prospective cohort imaging studies (German National Cohort Consortium, 2014; Breteler et al., 2014; Schram et al., 2014; Miller et al., 2016; Alfaro-Almagro et al., 2018) contain only a modest number of subjects with specific diseases (e.g., of the 100,000 participants projected to be scanned in the UK Biobank Alfaro-Almagro et al., 2018, only around 200 and 1000 can be expected to be multiple sclerosis and epilepsy patients, respectively Mackenzie et al., 2014; Joint Epilepsy Council, 2011).

As we shall see, it is straightforward to generalize the model of Fig. 1(a) to additionally include known subject-level covariates **y** that cause their own anatomical changes in **t** (independent of x). The resulting model, which is illustrated in Fig. 1(b), can be used to try and help improve prediction accuracy. When covariates are (partial) causes of x itself, as illustrated in Fig. 1(c), they are known as *confounders* – common underlying causes of both x and **t**. It is well-known (Pearl and Mackenzie, 2018) that confounders should be included in the model to preserve its ability to capture the causal effect of x on **t**, as will be illustrated in Section 5.1.

2.2. Scenarios not considered

Scenarios other than those shown in Fig. 1(a)–(c) are not considered in this paper. For instance, cases where confounders are not observed

[1] Assuming the absence of uncontrolled confounding variables, see Fig. 1.

(Fig. 1(d)) would bias the estimated causal effect of x on $\mathbf{t}$ to also include the effect of $\mathbf{y}$ on $\mathbf{t}$, which would lead to misleading visual explanations (see Section 5.1 for an example). We also do not consider *decoding* models (discriminative models, illustrated in Fig. 1(e)) in which the direction of causality is reversed, i.e., where the brain changes visualized in $\mathbf{t}$ are the *cause* (rather than the *effect*) of x. In experimental neuroscience, such models are used to investigate the mapping from functional or structural anatomy to perceptual and behavioral consequences or their deficits (Friston et al., 2008; Pölsterl and Wachinger, 2021; Chevalier et al., 2021). In these applications, the focus is on elucidating the mapping between $\mathbf{t}$ and x rather than on subject-level prediction, as x is already known.

2.3. Related work

The method we propose can be viewed as a generalization of naive Bayesian classifiers, in which the strong conditional independence assumption between input features is relaxed: When the number of latent variables is artificially clamped to zero in our method, the resulting predictor will devolve into a "naive" one. Naive methods have previously been shown to have surprisingly strong prediction performance in scenarios where the size of the training set is limited (Domingos and Pazzani, 1997; Ng and Jordan, 2002), in part because their simple structure prevents overfitting (Domingos and Pazzani, 1997; Domingos, 2012). Our findings indicate that this property also holds for the proposed method: In training regimes with up to a few thousand subjects – the typical scenario in many applications – we obtain prediction accuracies that rival those of the best image-based prediction methods available to date.

In its most basic (linear) form, the proposed method generates two spatial maps: a generative one that is suitable for human interpretation, and a discriminative one, computed from the generative one, that the method uses to make predictions. The distinct role in interpretation vs. prediction of these two different type of maps has been recognized before in the literature. In Haufe et al. (2014), for instance, the authors proposed a technique for computing a linear generative map that is compatible with a discriminative one and that is advocated to be more interpretable. However, as we demonstrate in Appendix D, this technique can be highly misleading as its always generates the same result irrespective of which image areas are actually used in the prediction computations, as long as the predictions themselves are accurate (see Fig. D.18). At the other end of the spectrum, in Varol et al. (2018) the authors developed a method in which the discriminative and the generative maps are forced to be identical. Although good performance was reported, the requirement to also make good predictions will inevitably bias the generative map, potentially limiting its validity as a tool for meaningful neuroanatomical interpretation.

Several methods exist that, like the proposed method, allow one to generate synthetic images simulating the effect of specific variables of interest on brain shape — either on a population level (e.g., age-specific brain templates Dalca et al., 2019; Pinaya et al., 2022; Wilms et al., 2022; Zhao et al., 2019) or for individual subjects (e.g., artificially aging an individual brain Pawlowski et al., 2020; Ravi et al., 2019; Wilms et al., 2022; Xia et al., 2021). However, only a few of these methods are designed to also be "inverted" to provide accurate single-subject predictions. Like our method, both (Zhao et al., 2019; Wilms et al., 2022) use latent variable models, but, unlike ours, they "decode" these latent variables using neural networks: Zhao et al. (2019) is based on a variational autoencoder (VAE), whereas Wilms et al. (2022) uses normalizing flows. The resulting nonlinearities increase the expressiveness of the models, but come at the price of additional computational complexity and the need for various approximations during training and inference (Zhao et al., 2019) or even of the input data itself (Wilms et al., 2022). In contrast, predicting, generating conditional templates, computing counterfactuals, and even training involve only evaluating analytical expressions with the proposed method. An experimental comparison of our method with the VAE of Zhao et al. (2019), detailed in Section 4.3, suggests that this simplicity does not come with a loss in prediction accuracy.

3. Basic linear version

In this section, we describe the proposed method in its most elementary form: a simple causal relationship between a variable of interest and the resulting image – the situation depicted in Fig. 1(a) – that is furthermore assumed to be linear. More complex models, with nonlinear dependencies on the variable of interest or the inclusion of subject-specific covariates and confounders, will be presented in Section 5.

3.1. Generative model

Let $\mathbf{t} \in \mathbb{R}^J$ denote a vector that contains the intensities in the J voxels of a subject's image, and x a scalar variable of interest about that subject (such as their age or gender). A simple generative model, illustrated in Fig. 2, is then of the form

$$\mathbf{t} = \mathbf{m} + x\mathbf{w}_G + \boldsymbol{\eta}. \tag{1}$$

Here $\mathbf{w}_G \in \mathbb{R}^J$ is a spatial weight map – referred to as the *generative* weight map in the remainder – that reflects how strongly the variable of interest x is expressed in the voxels of $\mathbf{t}$: Assuming a causal relationship between x and $\mathbf{t}$, it encodes how a unit increase x changes each voxel's intensity, on average. Further, $\mathbf{m} \in \mathbb{R}^J$ is a spatial template of intensities at baseline (i.e., when $x = 0$), and $\boldsymbol{\eta} \in \mathbb{R}^J$ is a random noise vector, assumed to be Gaussian distributed with zero mean and covariance $\mathbf{C}$. For notational convenience we will collect the two spatial weight maps $\mathbf{m}$ and $\mathbf{w}_G$ in a single matrix $\mathbf{W} = (\mathbf{m}, \mathbf{w}_G)$ for the remainder of the paper.

Note that this is the model commonly assumed in traditional mass-univariate brain mapping techniques, such as voxel- and deformation-based morphometry (Ashburner and Friston, 2000; Chung et al., 2001), where diagonal $\mathbf{C}$ is assumed and $\mathbf{w}_G$ is analyzed with statistical tests to reveal brain regions with significant effects. In contrast, here we assume that $\mathbf{C}$ has spatial structure, allowing us, besides interpreting $\mathbf{w}_G$, to accurately predict x from $\mathbf{t}$ by inverting the model, as shown below.

3.2. Making predictions

When the parameters of the model ($\mathbf{W}$ and $\mathbf{C}$) are known, the unknown target variable x^* of a subject with image $\mathbf{t}^*$ can be inferred by inverting the model using Bayes' rule. For a binary target variable $x^* \in \{0, 1\}$ with prior probability $p(x^*)$, it is well-known that the target posterior distribution takes the form of a logistic regression classifier (Hart et al., 2000) as shown in Appendix A:

$$p(x^* = 1|\mathbf{t}^*, \mathbf{W}, \mathbf{C}) = \sigma\left(\mathbf{w}_D^T\mathbf{t}^* + w_o\right), \tag{2}$$

where

$$\mathbf{w}_D = \mathbf{C}^{-1}\mathbf{w}_G \tag{3}$$

are a set *discriminative* spatial weights, $\sigma(a) = 1/(1 + e^{-a})$ denotes the logistic function, and $w_o = -\mathbf{w}_D^T(\mathbf{m} + \mathbf{w}_G/2) + \log\left[p(x^*=1)/p(x^*=0)\right]$. The maximum a posteriori (MAP) estimate of x^* is therefore 1 if

$$\mathbf{w}_D^T\mathbf{t}^* + w_o > 0, \tag{4}$$

and 0 otherwise. In the remainder of the paper, we will assume equal priors: $p(x^*=1) = p(x^*=0) = 0.5$ unless stated otherwise.

For a continuous target variable with a flat prior $p(x^*) \propto 1$, the posterior distribution is Gaussian with variance

$$v = \left(\mathbf{w}_G^T\mathbf{C}^{-1}\mathbf{w}_G\right)^{-1} \tag{5}$$

and mean

$$\tilde{x}^* = v(\mathbf{w}_D^T\mathbf{t}^* + b_0), \tag{6}$$

where $b_0 = -\mathbf{w}_D^T\mathbf{m}$ (see Appendix A). The predicted value of x^* is therefore given by (6), which again involves taking the inner product of the discriminative weights $\mathbf{w}_D$ with $\mathbf{t}^*$. An example of model inversion in case of age prediction is shown in Fig. 3.

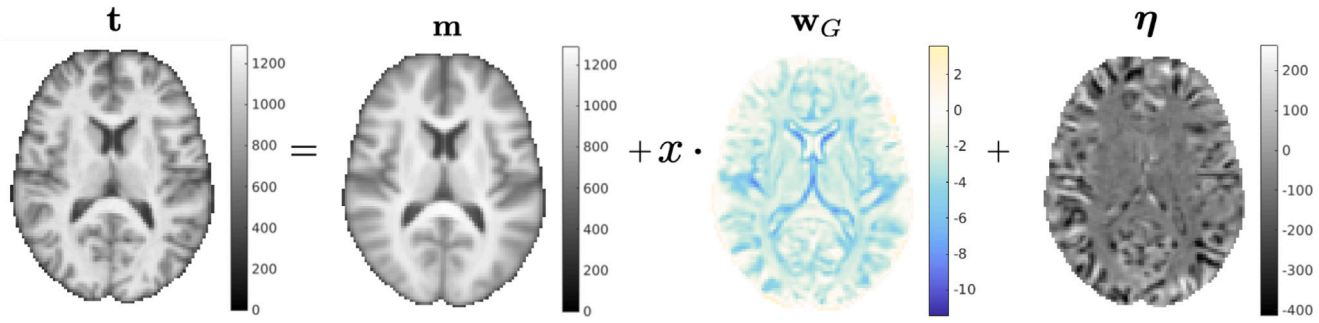


Fig. 2. Example of the forward model (1), applied to modeling the effect of age on brain morphology. Here x denotes the difference between the age of the subject and the average age in a training set.

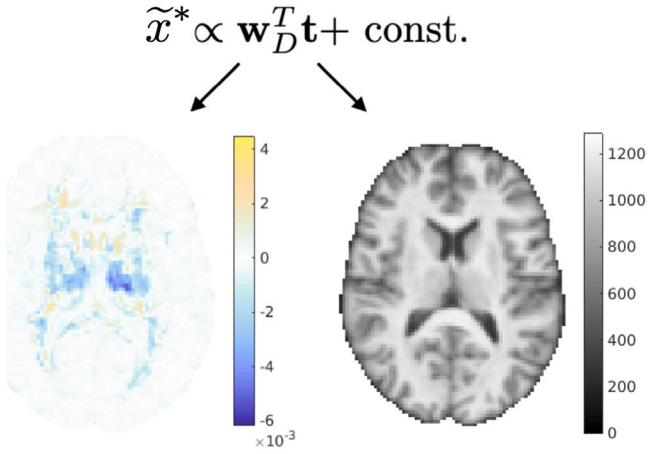


Fig. 3. Illustration of how a subject's age is estimated by inverting the model shown in Fig. 2.

3.3. Model training

In practice the model parameters $\mathbf{W}$ and $\mathbf{C}$ need to be estimated from training data. Given N training pairs $\{\mathbf{t}_n, x_n\}_{n=1}^N$, their maximum likelihood (ML) estimate is obtained by maximizing the marginal likelihood

$$p\left(\{\mathbf{t}_n\}_{n=1}^N | \{x_n\}_{n=1}^N, \mathbf{W}, \mathbf{C}\right) = \prod_{n=1}^N \mathcal{N}\left(\mathbf{t}_n | \mathbf{W}\boldsymbol{\phi}_n, \mathbf{C}\right) \tag{7}$$

with respect to $\mathbf{W}$ and $\mathbf{C}$, where we have defined $\boldsymbol{\phi}_n = (1, x_n)^T$. For the spatial maps, the solution is given in closed form (see Appendix B):

$$\mathbf{W} = \left(\sum_{n=1}^N \mathbf{t}_n \boldsymbol{\phi}_n^T\right)\left(\sum_{n=1}^N \boldsymbol{\phi}_n \boldsymbol{\phi}_n^T\right)^{-1}. \tag{8}$$

However, obtaining the noise covariance matrix $\mathbf{C}$ directly by ML estimation is problematic: $\mathbf{C}$ has $J(J+1)/2$ free parameters, which is orders of magnitude more than there are training samples (recall that J is the number of voxels). To be able to control the number of parameters while still capturing the dominant correlations in the noise, we impose a specific structure on $\mathbf{C}$ by using a latent variable model known as factor analysis (Bishop, 2006). In particular, we model the noise as

$$\boldsymbol{\eta} = \mathbf{V}\mathbf{z} + \boldsymbol{\epsilon}, \tag{9}$$

where $\mathbf{z}$ is a small set of K unknown latent variables distributed as $p(\mathbf{z}) = \mathcal{N}(\mathbf{z}|\mathbf{0}, \mathbb{I}_K)$, $\mathbf{V}$ contains K corresponding, unknown spatial weight maps, and $\boldsymbol{\epsilon}$ is a zero-mean Gaussian distributed error with unknown diagonal covariance $\boldsymbol{\Delta}$. Marginalizing over $\mathbf{z}$ yields a zero-mean Gaussian noise model (Bishop, 2006) with covariance matrix

$$\mathbf{C} = \mathbf{V}\mathbf{V}^T + \boldsymbol{\Delta},$$

which is now controlled by a reduced set of parameters $\mathbf{V}$ and $\boldsymbol{\Delta}$. The number of columns in $\mathbf{V}$ (i.e., the number of latent variables K) is a hyperparameter in the model that needs to be tuned experimentally.

Plugging in the ML estimate of $\mathbf{W}$ given by (8), the parameters $\mathbf{V}$ and $\boldsymbol{\Delta}$ maximizing the marginal likelihood (7) can be estimated using an Expectation–Maximization (EM) algorithm (Rubin and Thayer, 1982). Defining

$$\boldsymbol{\eta}_n = \mathbf{t}_n - \mathbf{W}\boldsymbol{\phi}_n \tag{10}$$

as the noise vector of training subject n, this yields an iterative algorithm that repeatedly evaluates the posterior distribution over the latent variables:

$$p(\mathbf{z}_n|\boldsymbol{\eta}_n, \mathbf{V}, \boldsymbol{\Delta}) = \mathcal{N}(\mathbf{z}_n|\boldsymbol{\mu}_n, \boldsymbol{\Sigma}), \quad \forall n \tag{11}$$

where $\boldsymbol{\mu}_n = \boldsymbol{\Sigma}\mathbf{V}^T\boldsymbol{\Delta}^{-1}\boldsymbol{\eta}_n$ and $\boldsymbol{\Sigma} = (\mathbb{I}_K + \mathbf{V}^T\boldsymbol{\Delta}^{-1}\mathbf{V})^{-1}$, and subsequently updates the parameters accordingly:

$$\mathbf{V} \leftarrow \left(\sum_{n=1}^N \boldsymbol{\eta}_n \boldsymbol{\mu}_n^T\right)\left(\sum_{n=1}^N \left(\boldsymbol{\mu}_n\boldsymbol{\mu}_n^T + \boldsymbol{\Sigma}\right)\right)^{-1} \tag{12}$$

$$\boldsymbol{\Delta} \leftarrow \operatorname{diag}\left(\frac{1}{N}\sum_{n=1}^N \boldsymbol{\eta}_n\boldsymbol{\eta}_n^T - \mathbf{V}\frac{1}{N}\sum_{n=1}^N \boldsymbol{\mu}_n\boldsymbol{\eta}_n^T\right). \tag{13}$$

Here $\operatorname{diag}(\cdot)$ sets all the non-diagonal entries to zero.

3.4. Practical implementation

The method outlined above involves manipulating matrices of size $J \times J$. Despite the high dimensionality (recall that J is the number of voxels), computations can be performed efficiently by exploiting the structure of these matrices: As detailed in Appendix C, both training and predicting can be implemented in a way that only involves the posterior covariance of the latent variables $\boldsymbol{\Sigma}$, which is of much smaller size $K \times K$.

In our implementation, we center the target variable x, i.e., we use values from which the sample mean in the training set has been subtracted. As shown in Appendix D, this has the advantage that the estimated template $\mathbf{m}$ then represents the anatomy of the "average" subject in the training set, i.e., $\mathbf{m} = 1/N \sum_{n=1}^N \mathbf{t}_n$. For estimating the parameters $\boldsymbol{V}$ and $\boldsymbol{\Delta}$ of the noise model, we first perform a voxel-wise rescaling of the noise vectors $\{\boldsymbol{\eta}_n\}_{n=1}^N$, such that each voxel has unit variance across the training subjects. We then initialize the EM algorithm by using a matrix with standard Gaussian random entries for $\mathbf{V}$, and the identity matrix for $\boldsymbol{\Delta}$. Convergence of the EM procedure is detected when the relative change in the log marginal likelihood drops below 10^{-5} between iterations. The elements in the estimated $\mathbf{V}$ and $\boldsymbol{\Delta}$ are then rescaled back to the original intensity space to obtain the final parameters of the noise model.

The code for the proposed model is available at https://github.com/chiara-mauri/Interpretable-subject-level-prediction, with both Matlab and Python implementations.

4. Experiments on age and gender prediction

In this section, we present experiments on the task of predicting a subject's age and gender from their brain MRI scan. Specifically, we compare the behavior of the basic linear model described in Section 3 with that of three state-of-the-art benchmark methods, when the size of the training set is varied.

4.1. Experimental set-up

The benchmarks we used consist of a nonlinear-discriminative (SFCN Peng et al., 2021), a linear-discriminative (RVoxM Sabuncu and Van Leemput, 2012), and a nonlinear-generative (VAE Zhao et al., 2019) method for image-based prediction. Together with the proposed method, which is linear-generative in the basic form analyzed here, these benchmarks form a representative sample of the spectrum of methods available to date.

We trained each of these methods on randomly sampled subsets of 26,127 T1-weighted scans of healthy subjects drawn from the UK Biobank (Alfaro-Almagro et al., 2018). Each of these subjects was between 44 and 82 years old, and was scanned on one of three identical 3T scanners using the same MRI protocol. In our experiments, we used the skull-stripped and bias-field corrected T1-weighted scans that are made publicly available, which are computed using a nonlinear registration with the MNI152 template (see Alfaro-Almagro et al. (2018) for details). We took advantage of these nonlinear registrations to warp and resample all the scans to the same MNI152 template space using linear interpolation. The resulting 1 mm isotropic T1-weighted images then formed the input of the various image-prediction algorithms.

To analyze the behavior of the different methods when the training set size is gradually increased from 100 to almost 10,000 subjects, we trained each method multiple times for each training set size. Specifically, each training run was repeated 10 times, with different randomly sampled training subjects, except for the larger training sizes ($N > 1000$) where the number of repetitions was limited to 3 to reduce the computational burden. After training, the prediction performance of each model was evaluated on a fixed set of 1000 randomly sampled test subjects not overlapping with the training subjects. The size of this test set (and of a separate validation set, see below) was chosen to be similar to the one used in Peng et al. (2021) so that our results could be compared to those reported in that paper. For the age prediction task, the Mean Absolute Error (MAE)[2] was used as the evaluation criterion, whereas the average classification accuracy was used in the gender prediction experiments.

Consistent with the set-up of Peng et al. (2021), we also had a separate, fixed validation set of 500 randomly sampled subjects. This was used to tweak the hyperparameter(s) for each method (see details below), using a grid search to optimize MAE (for age) and classification accuracy (for gender) for each training run. The details about the hyperparameter values considered in the grid search for each method are provided in Sec. 3 of the supplementary material.

4.2. Implementation of the benchmark methods

The four methods under comparison were implemented as follows:

- **Proposed method**: We performed all the experiments with the proposed method in Matlab, running on a Linux CPU machine (Intel Xeon E5-2660V3 10 Core CPU 2.60 GHz, 128 GB RAM). To speed up computations, the number of voxels was reduced by masking out the background (by thresholding the average of all images in each training set), and by downsampling the input data to a 3 mm isotropic resolution. This downsampling was found not to affect the prediction performance in pilot experiments (see supplementary material). For each training run, performance on the validation set was used to set the number of latent variables K, which is the only hyperparameter of the method. An example of a model trained this way is shown in Figs. 4–6.
- **RVoxM**: This is a discriminative method that imposes sparsity and spatial smoothness on its weight map as a form of regularization (Sabuncu and Van Leemput, 2012). As reported in the supplementary material, it yields competitive prediction performances compared to other commonly used linear-discriminative methods, and was therefore chosen to represent this family of models (although other methods could also have been chosen). We used the Matlab code that is publicly available,[3] with some adaptations to make it more efficient (by parallelizing part of the training loop) and, for gender prediction, more robust to high-dimensional input. The same background masking, downsampling procedure and computer hardware was used as for the proposed method in the experiments. The method has a single hyperparameter to control the spatial smoothness of the weight map that it computes, which was tuned on the validation set for each training run.
- **SFCN**: This is the lightweight convolutional neural network proposed by Peng et al. (2021), who, to the best of our knowledge, have reported the best performance for brain age prediction to date. Since code for training this method is not publicly available, we modified the implementation of Mouches et al. (2022)[4] to match the description provided in Peng et al. (2021) as closely as possible. In particular, we replicated the same data augmentation scheme, network architecture, L2 weight decay and batch size, deviating only in the last network layer as discretizing age into 40 bins (as described in Peng et al. (2021)) did not benefit prediction accuracy in our experiments. The method directly takes 1 mm isotropic images as input, and was therefore run on a high-end GPU with a sufficiently large amount of memory (NVIDIA A100 SXM4 GPU with 40 GB of RAM) in our experiments. For each training run, the validation set was used to determine the optimal number of training epochs of this method.
- **VAE**: This is a generative method for age prediction (Zhao et al., 2019) with publicly available training code.[5] It can be regarded as a generalization of the proposed method, where the latent variables are expanded *nonlinearly* through a deep neural network, which precludes exact model inversions and makes computations more involved. Since the method is designed to work with images that are downsampled to 2 mm isotropic resolution and cropped around the ventricles (Zhao et al., 2019), the comparison with this method was performed separately from the other two benchmarks, with both the VAE and the proposed method running on the same cropped 2 mm volumes. We tested only training sizes in a similar range (from 100 to 400 training subjects) as the one used in the original paper (196 subjects), running the VAE on a RTX 6000 GPU with 65G of RAM. Although the method originally only contains two hyperparameters (dropout factor and L2 regularization), for a fair comparison we also varied the number of latent variables (originally hardcoded to 16) as an extra hyperparameter. As for the other methods, the values of the three resulting hyperparameters were tuned on the validation set for each training run.

[2] MAE is defined as the absolute difference between predicted and real value of the target variable, averaged across all test subjects.

[3] https://sabuncu.engineering.cornell.edu/software-projects/relevance-voxel-machine-rvoxm-code-release/.

[4] https://github.com/pmouches/Multi-modal-biological-brain-age-prediction/blob/main/sfcn_model.py.

[5] https://github.com/QingyuZhao/VAE-for-Regression.

Fig. 4. The estimated template **m** (the average image in the training set) when the model is trained on $N = 2600$ subjects in an age prediction task.

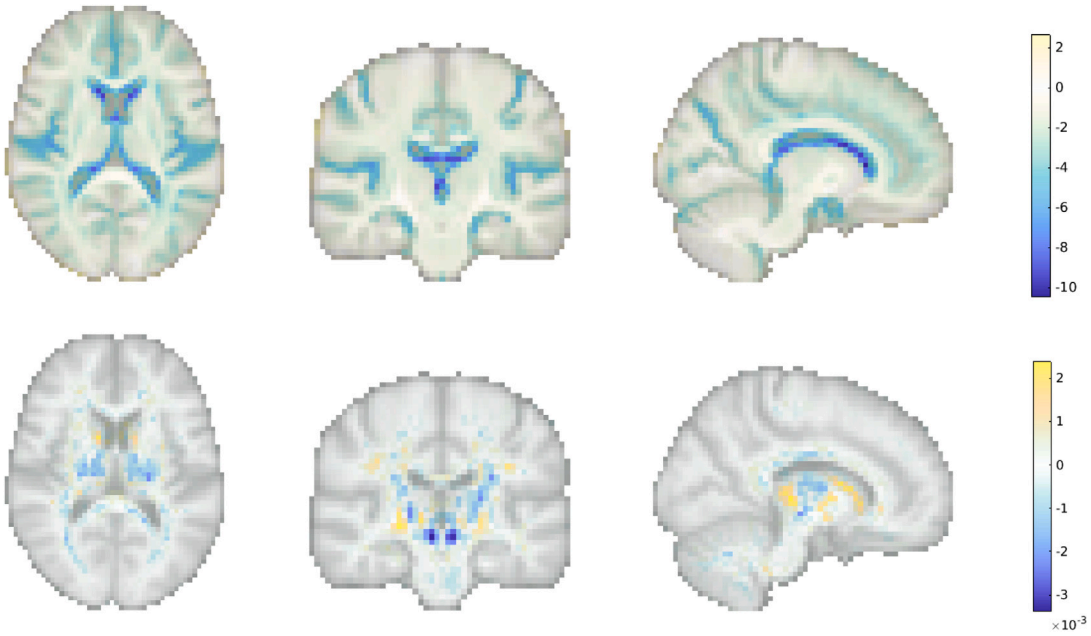


Fig. 5. Top: the generative map $\mathbf{w}_G$ – expressing the effect of aging – estimated from $N = 2600$ subjects, overlaid on the template of Fig. 4. Voxels with zero weight are transparent. Bottom: the corresponding discriminative map $\mathbf{w}_D$ that is used for making age predictions. The discrepancy between these two maps is analyzed in Section 4.5.

4.3. Prediction performance and training times

Fig. 7 shows the prediction performance of the proposed method, RVoxM and SFCN as the number of training subjects is varied, both for predicting age (Fig. 7 left) and gender (Fig. 7 right). As expected, all prediction performances improve as the training size increases. In the age prediction task, the proposed method yields generally the best performance when the training set size is up to 2600 subjects, after which SFCN is the better method. For gender prediction, SFCN is clearly outperformed by both RVoxM and the proposed method.

For completeness, Fig. 7 also includes the results for SFCN as reported by its authors in Peng et al. (2021), where a similar experimental set-up as ours was used. It can be seen that, although our implementation closely followed the description provided by the authors, we were not always able to match their reported performance: Especially for gender prediction, and for age prediction with very large training sets, there are considerable discrepancies between the two implementations. One possible explanation is that SFCN is a complex model with many more "knobs" to be tuned correctly than the proposed method, which has only a single hyperparameter. Another explanation is that the experimental set-up is not entirely comparable: In Peng et al. (2021) the model is only trained on one training set for each training size; the subjects in their test set are different from ours; and they used affinely instead of nonlinearly registered scans (although the latter point is reported to yield only minimal differences in Peng et al. (2021), which we can confirm based on our own experiments).

Table 1 reports, for the age prediction experiment, the training times required by the proposed method, RVoxM and SFCN. SFCN is by far the slowest to train, requiring many hours even for very small training sizes, and several days for large ones. Up to training sizes of 1000 subjects, the proposed method is considerably faster to train than RVoxM (dozens of minutes CPU time at most), but slows down for larger training sizes. This can be explained by the fact that the optimal number of latent variables in our model (hyperparameter K) increases rapidly with the number of training subjects, as shown in Table 2. When interpreting the training times of RVoxM and the proposed method on the one hand, and those of SFCN on the other, it should be taken into account that the latter uses hardware acceleration but also works with much larger (non-downsampled) input volumes.

Finally, Fig. 8 shows the age prediction performance of the VAE and the proposed method, as a function of the training set size. The proposed method achieves better results for every training size tested. This suggests that, at least when only a few hundred subjects are available for training, adding nonlinearities in the method's noise model is not beneficial. Furthermore, the VAE is considerably slower to train than the proposed method: around 8 min GPU time for 200 training subjects, vs. 1 min CPU time with the proposed method.

4.4. Bias–variance trade-off

More insight into the prediction performances reported in the previous section can be obtained using the so-called bias–variance decomposition. Specifically for age, which is a continuous variable, a

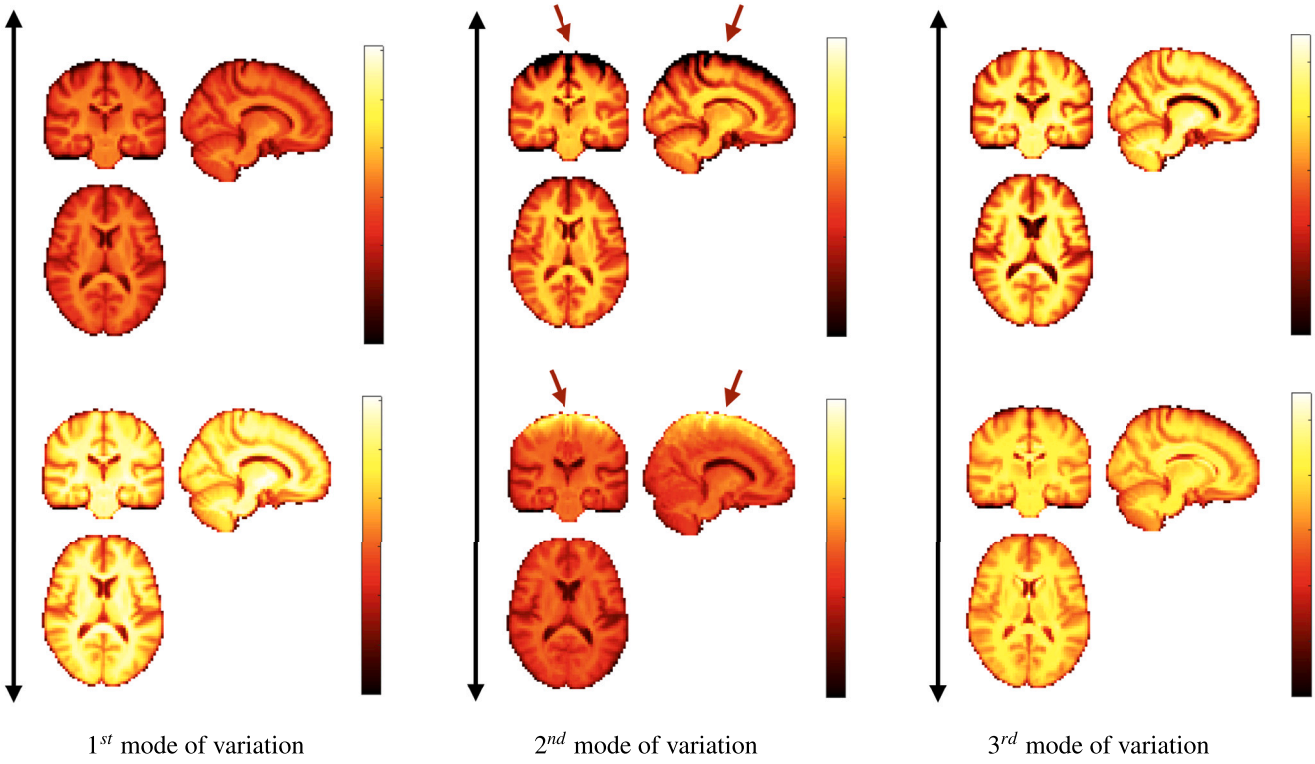


Fig. 6. Illustration of the major modes of variation captured by the linear-Gaussian noise model, when trained on $N = 2600$ subjects in an age prediction task. Each mode illustrates the effect of applying one of the three first eigenvectors of $\mathbf{V}\mathbf{V}^T$ (the component of $\mathbf{C}$ that has spatial structure) on the template of Fig. 4. (We note that directly showing the learned basis vectors in the columns of $\mathbf{V}$ is not meaningful, since the model is invariant to rotations in the latent space (Bishop, 2006): using $\tilde{\mathbf{V}} = \mathbf{V}\mathbf{R}$ for an arbitrary orthogonal matrix $\mathbf{R}$ yields the same covariance matrix $\tilde{\mathbf{V}}\tilde{\mathbf{V}}^T = \mathbf{V}\mathbf{V}^T$.) The top and bottom figures show the template modified in the positive and negative direction of the eigenvectors, respectively. Note that the first eigenvector seems to encode a general brightening/darkening of the image intensities, while the second one models residual bias fields that were not removed in the preprocessing of the UK Biobank data.

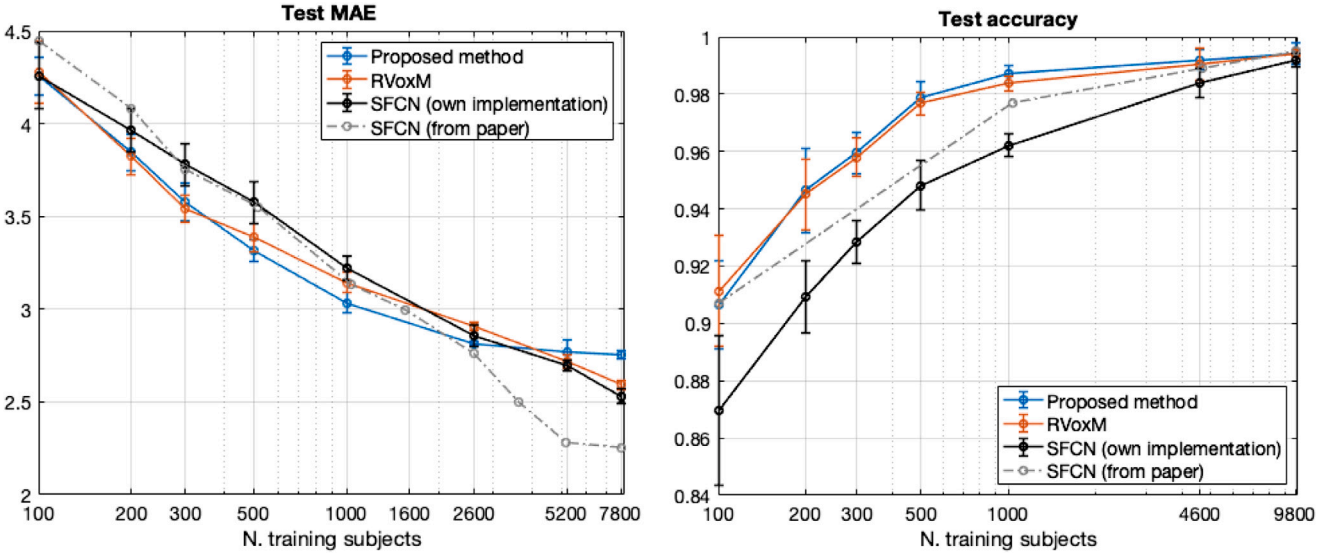


Fig. 7. Prediction performance of the proposed method, RVoxM and SFCN for age (left) and gender (right) as the number of training subjects is varied. For each method, the full line shows the average performance across multiple training runs, and the whiskers extend to one standard deviation away from the average. The test MAE for age is indicated in years. For reference, we also include the performance of SFCN as reported in Peng et al. (2021), although the results are not entirely comparable.

Table 1
Training times in the age prediction task for the proposed method, RVoxM and SFCN. These training times were evaluated at the optimal value of a single hyperparameter for each method, determined on an external validation set. The reported times are the average across all training runs for each training set size N.

	N = 100	N = 200	N = 300	N = 500	N = 1000	N = 2600	N = 5200	N = 7800
Proposed method	1.20 min	0.67 min	1.94 min	9.53 min	32 min	≈3 h	≈15 h	≈69 h
RVoxM	92 min	66 min	75 min	76 min	129 min	127 min	≈22 h	≈21 h
SFCN	≈8 h	≈11 h	≈16 h	≈18 h	≈34 h	≈76 h	≈69 h	≈102 h

Table 2
Optimal number of latent variables K of the proposed method when predicting age, as determined on an external validation set. The reported number is the rounded average across all training runs for each training set size N.

	N = 100	N = 200	N = 300	N = 500	N = 1000	N = 2600	N = 5200	N = 7800
K	20	21	52	86	120	367	1833	3333

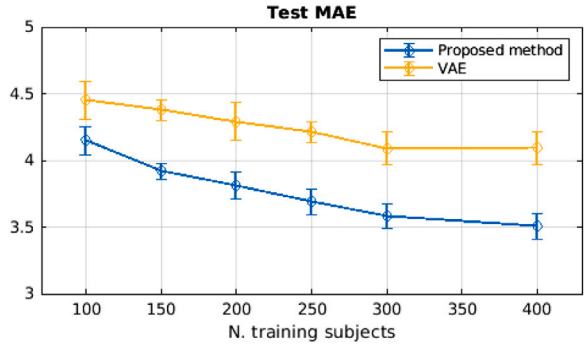


Fig. 8. Age prediction performance of the VAE and the proposed method, when applied to 2 mm isotropic images cropped around the ventricles, represented in the same way as in Fig. 7.

particular method's mean squared error (MSE) can be decomposed as follows (Bishop, 2006):

$$\underbrace{\mathbb{E}_D\left[\left(x^* - \tilde{x}_D^*\right)^2\right]}_{MSE} = \underbrace{\left(x^* - \mathbb{E}_D\left[\tilde{x}_D^*\right]\right)^2}_{bias^2} + \underbrace{\mathbb{E}_D\left[\left(\tilde{x}_D^* - \mathbb{E}_D\left[\tilde{x}_D^*\right]\right)^2\right]}_{variance}. \quad (14)$$

Here x^* denotes the real age of a given test subject, $\tilde{x}_D^*$ is the predicted age when the method is trained on a particular dataset D of a certain size, and $\mathbb{E}_D[\cdot]$ denotes the average over multiple such training sets. In our set-up, in which each method is trained multiple times using different randomly sampled subjects, the *bias* component in (14) reflects a systematic error that is persistent across the training runs, whereas the *variance* component indicates how much the predictions change between the different training runs (Hart et al., 2000; Bishop, 2006; Domingos and Pazzani, 1997). Typically, flexible models tend to have low bias but high variance, reflecting an *overfitting* to the training data, while strongly constrained methods display the opposite behavior, resulting in *underfitting* of the training data (Hart et al., 2000; Bishop, 2006; Domingos and Pazzani, 1997).

Fig. 9 (left) shows how the bias, the variance and the resulting MSE change in the age prediction experiment, when the training set size is varied. The various curves were obtained by averaging (14) across all test subjects for our method (blue), RVoXM (red) and SFCN (black). The proposed method generally has the highest bias among the three methods; however this is off-set by a lower variance, resulting in a strong overall prediction performance in training set sizes of up to 2600 subjects. Its low variance is obtained by controlling the flexibility of the method: when the number of training subjects is small, only a limited number of latent variables is selected (see Table 2), resulting in a simple, highly regularized model that successfully avoids overfitting to the training data. As the training size increases, the number of latent variables is allowed to grow, resulting in gradually more flexible models with decreased bias and therefore better prediction performance (see Fig. 9 (left)). However, for very large training sets (over 2600 subjects), the method's strong modeling assumptions prevent it from decreasing its bias further, and the nonlinear SFCN can now take advantage of its flexibility (lower bias) without overfitting (variance comparable to the proposed method), resulting in a better prediction performance.

Fig. 9 (right) shows the results of the bias–variance decomposition for the VAE and our method, when applied to age prediction from 2 mm cropped data. It can be seen that the VAE's lower prediction performance (reported in Section 4.3) can be attributed to its significantly higher bias, presumably due to the variational approximations (Kingma and Welling, 2013) that are used to invert its model.

4.5. Interpretability analysis

A key advantage of the proposed method over discriminative methods such as RVoxM and SFCN is that, in addition to the discriminative map $\mathbf{w}_D$ that it uses to make predictions, it also computes a generative map $\mathbf{w}_G$ that expresses the causal effect of the variable of interest on brain morphology. To illustrate why this is important, Fig. 10 shows, for three different training set sizes, the discriminative map $\mathbf{w}_D$ computed by our method for predicting age, along with the corresponding discriminative map of RVoXM and the SmoothGrad saliency map (Smilkov et al., 2017) – a generalization of linear spatial maps to nonlinear methods (Adebayo et al., 2018) – of SFCN. The inconsistencies of these maps across both the training set sizes and the different methods, and their overall lack of correspondence with the known neurobiology of aging, illustrate the difficulty of using discriminative maps for human interpretation.

More insight can be gained by examining the proposed method specifically, since it uses discriminative maps that are derived from generative ones. It is worth noting that estimating the generative maps from training data is itself quite stable, since it merely amounts to fitting two basis functions to hundreds of measurements in each voxel (see (8)). Furthermore, as illustrated in Fig. 11, the resulting maps are intuitive to interpret, since they show typical age-related effects such as cortical thinning and ventricle enlargement (Fjell et al., 2009; Fjell and Walhovd, 2010). When the discriminative maps are subsequently computed as $\mathbf{w}_D = \mathbf{C}^{-1}\mathbf{w}_G$, however, a strong dependency on the training set size is introduced, because the method explicitly controls the complexity of its noise model $\mathbf{C}$ in response to the size of the available training set (the bias–variance trade-off of Section 4.4). $\mathbf{C}$ can also capture peculiarities in the data that may be relevant for improving prediction performance, but not for human interpretation. An example of this was shown in Fig. 6, where overall brightness variations and residual MR bias field artifacts were picked up by the noise model. Through $\mathbf{C}$, such noise patterns can find their way into $\mathbf{w}_D$, producing hard-to-interpret spatial maps that no longer reflect the expected age-related brain atrophy patterns. This is clearly illustrated in Fig. 5,

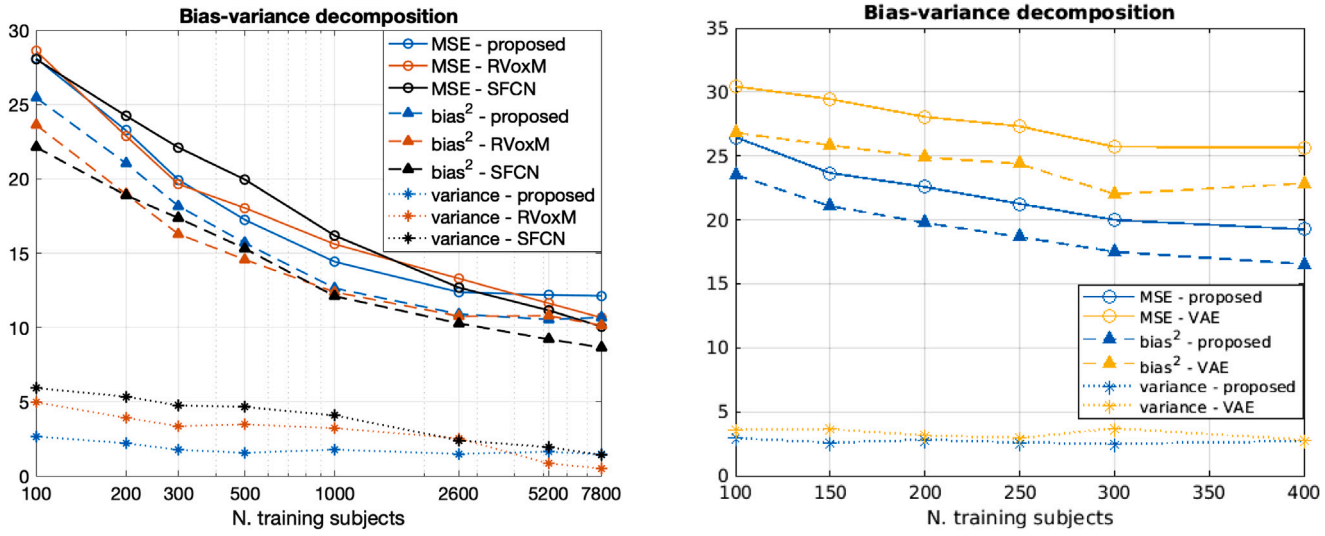


Fig. 9. Left: bias–variance decomposition for the age prediction results shown in Fig. 7 (left). Right: the same for the results shown in Fig. 8.

where the discriminative map $\mathbf{w}_D$ is contrasted with the corresponding generative map $\mathbf{w}_G$.

The reason the noise covariance $\mathbf{C}$ is taken into account in $\mathbf{w}_D$ – which in turn makes $\mathbf{w}_D$ hard to interpret – can be illustrated with a simple toy example involving only two "voxels", shown in Fig. 12. When the method is tasked with computing the prediction $\tilde{x}^*$ from an image $\mathbf{t}^*$, it effectively decomposes $\mathbf{t}^*$ into its most likely constituent components: $\mathbf{m}$ (a population template), $\tilde{x}^*\mathbf{w}_G$ (the estimated causal effect), and $\tilde{\boldsymbol{\eta}}^*$ (the most likely noise vector) – see (1) and Fig. 2. In this process, parts of the signal that are well-explained by the noise model $\mathbf{C}$ can be attributed to the noise component $\tilde{\boldsymbol{\eta}}^*$, and therefore effectively discarded when estimating $\tilde{x}^*$. This is illustrated in Fig. 12 (middle). Mathematically, the same value $\tilde{x}^*$ can also be obtained by simply computing the inner product $\mathbf{w}_D^T\mathbf{t}^*$ (see (6) and Fig. 3), which folds the process of separating the effect of interest from noise patterns into a single operation. This amounts to projecting the given data point $\mathbf{t}^*$ orthogonally onto the direction $\mathbf{w}_D$, as illustrated in Fig. 12 (right). However, this projection operation is hard to interpret on its own: $\mathbf{w}_D$ only reveals the final recipe of how predictions are computed, but no longer the underlying logic.

In addition to visualizing its generative maps directly for human interpretation, the causal interpretation of our model can also be used to simulate the effect of aging on actual images. On the population level, this can be achieved by computing age-specific templates $\mathbf{m} + x\mathbf{w}_G$ for different values of x, as illustrated in Fig. 13. More detailed explanations can also be provided at the level of the individual subject, using *counterfactuals* (Pearl and Mackenzie, 2018) – imaginary images of specific individuals if they had been younger or older. Given an image $\mathbf{t}$ and the real age x, (10) can be used to compute the subject-specific noise vector $\boldsymbol{\eta}$, which captures the subject's individual idiosyncrasies that are not explained by the population-level causal model. Counterfactual images can then be obtained by re-assembling the forward model from its constituent components, using a different, imaginary age x in (1). Examples of this process are shown in Fig. 14.

The ability to generate synthetic images that are conditioned on the target variable x provides alternative visual explanations of the method's prediction process beside simply showing the generative map $\mathbf{w}_G$. These will be particularly useful when the model is extended to include nonlinear effects (so that the cause–effect relationship can no longer be described using a single linear map, see Section 5.2), or when features are used that are more difficult to interpret than the voxel-level intensities used here (e.g., parameters of a deformation field, as discussed in Section 6). Counterfactual images in particular are thought to be closely aligned with human intuition (Pearl and Mackenzie, 2018): In a disease classification task where a subject is predicted to suffer from a particular disease, for instance, a counterfactual would show their presumed brain shape in health for comparison, mimicking how a human expert would explain their prediction to others.

5. More complex variants

Here we show how the basic form of the model described in Section 3 can easily be extended to take into account additional subject-specific covariates and confounders, as well as nonlinear dependencies on the variable of interest.

5.1. Covariates and confounders

When additional demographic or clinical information is available about the subjects, it is straightforward to include this in the model — either as additional covariates (Fig. 1(b)) or as confounders (Fig. 1(c)). This can be useful for further improving prediction accuracy, or for removing the effect of confounders that would otherwise invalidate the causal interpretation of the generative maps as demonstrated below.

Assuming each subject has L extra variables $y^1, \dots, y^L$, the model (1) can be extended to

$$\mathbf{t} = \mathbf{m} + x\mathbf{w}_G + \sum_{l=1}^{L} y^l \mathbf{w}_y^l + \boldsymbol{\eta},$$

where $\mathbf{w}_y^1, \dots, \mathbf{w}_y^L$ are now extra spatial weight maps that also need to be estimated. During training, the corresponding parameters $\mathbf{W} = (\mathbf{m}, \mathbf{w}_G, \mathbf{w}_y^1, \dots, \mathbf{w}_y^L)$, $\mathbf{V}$ and $\boldsymbol{\Delta}$ can still be estimated using (8), (12) and (13), provided that $\boldsymbol{\phi}_n = (1, x_n, y_n^1, \dots, y_n^L)^T$ is used instead of $\boldsymbol{\phi}_n = (1, x_n)^T$. To predict an unknown variable of interest x^* from a subject with image $\mathbf{t}^*$ and covariates $y^{*1}, \dots, y^{*L}$, (4) and (6) remain valid when $\mathbf{t}^*$ is replaced by $\left(\mathbf{t}^* - \sum_l y^{*l}\mathbf{w}_y^l\right)$.

To demonstrate the resulting model, we considered a classification experiment of Alzheimer's disease (AD) patients vs. healthy controls. For this experiment, we used the OASIS-1 dataset[6] (age range 18–96 years) consisting of T1-weighted scans of 100 AD subjects (average age 77 ± 7.12 years) and 336 controls (average age 44 ± 24 years). Importantly, we included age as a known confounder, since it both affects the imaging data $\mathbf{t}$ directly and is a partial cause of the disease status x (because AD occurs more frequently in older subjects, and the

[6] https://www.oasis-brains.org.

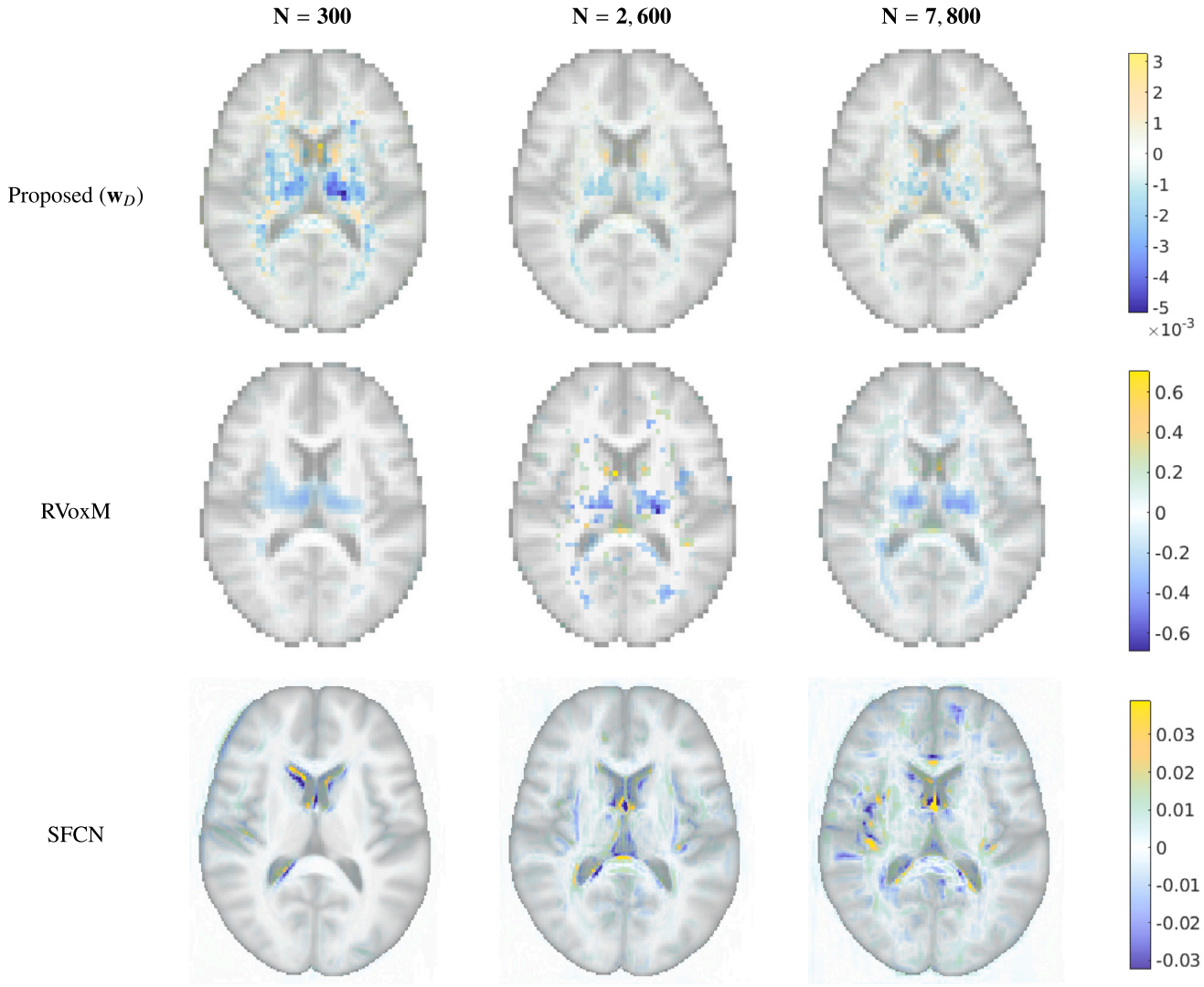


Fig. 10. Discriminative maps of proposed method, RVoxM, and SFCN (with SmoothGrad) obtained for age prediction with varying training set sizes, overlaid on a template. Voxels with zero weight are transparent. Since SmoothGrad yields subject-specific maps, we averaged across all test subjects to obtain population-level maps for SFCN.

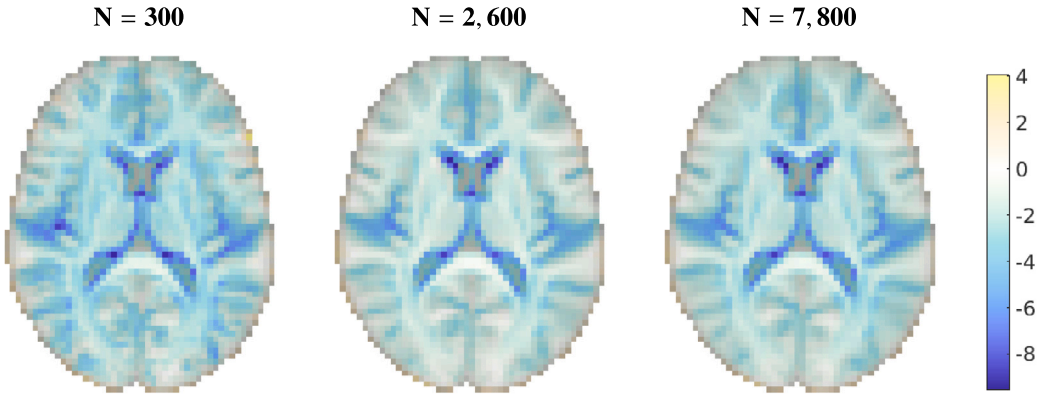


Fig. 11. Generative maps $\mathbf{w}_G$ computed by the proposed method for age prediction, for different training set sizes. In addition to expressing known aging patterns in the brain, these maps also show consistency across the different training sizes.

two disease groups are not age-matched in this dataset). Because there are also more controls than AD subjects, we used unequal priors of 0.77 vs. 0.23, respectively.

As input features we used probabilistic gray matter segmentations computed with SPM12, after warping them to standard space and modulating them (Ashburner et al., 2014) to preserve signal that would otherwise be removed by the spatial normalization. Given the limited number of subjects in the dataset, we determined the optimal value of the method's hyperparameter K in a nested 5-fold cross-validation (CV) setting: For each data split of the outer CV loop into a training and a

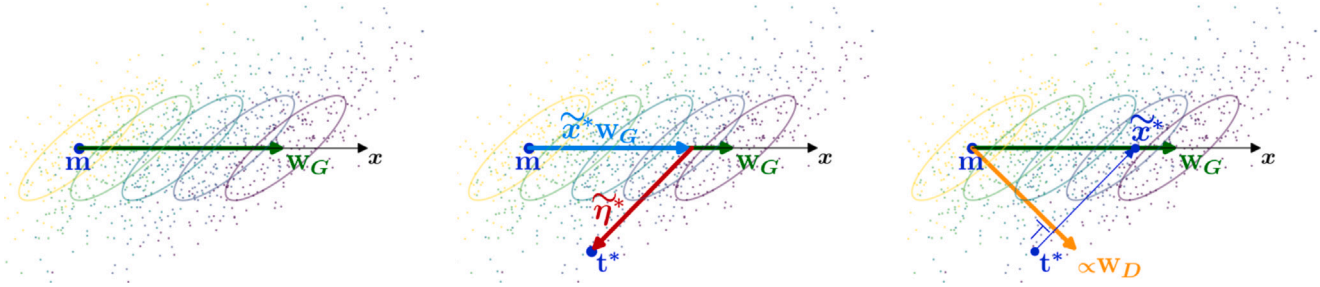


Fig. 12. Left: illustration of a the image generation process in a toy example. The figure shows the template $\mathbf{m}$, the generative map $\mathbf{w}_G$, and contour plots of the noise covariance matrix $\mathbf{C}$. Also shown are randomly generated 2D data points (dots in the figure), corresponding to 5 discrete values of a continuous variable of interest x (marked with different colors). The discretization was performed for visualization purposes. Middle: when the model is inverted to compute a predicted value $\tilde{x}^*$ for a given data point $\mathbf{t}^*$, the signal is effectively decomposed into $\mathbf{t}^* = \mathbf{m} + \tilde{x}^*\mathbf{w}_G + \tilde{\boldsymbol{\eta}}^*$. Right: the same result $\tilde{x}^*$ can also be obtained by projecting the data point $\mathbf{t}^*$ orthogonally onto the direction $\mathbf{w}_D = \mathbf{C}^{-1}\mathbf{w}_G$. This operation is *mathematically* equivalent to decomposing the signal into its components as in the middle figure, but not *in terms of interpretability*, since causal effect and noise pattern are intermingled in $\mathbf{w}_D$.

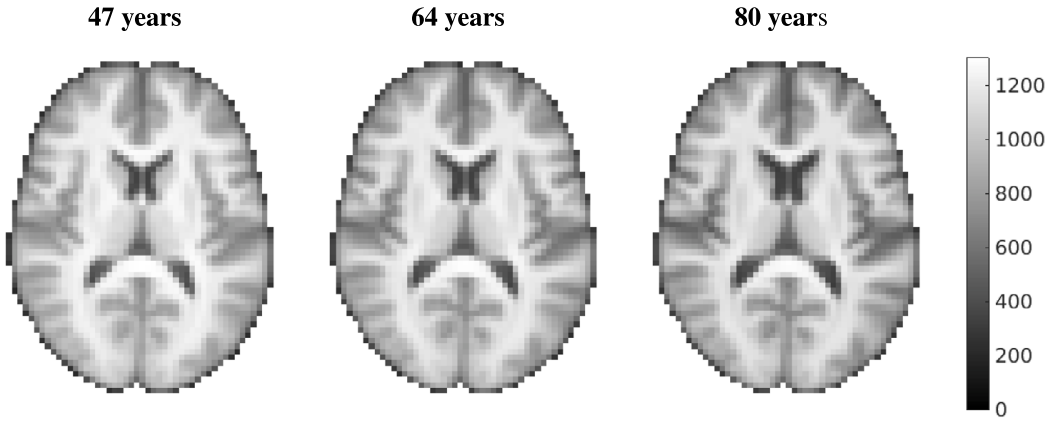


Fig. 13. Age-specific templates synthesized by the proposed method trained on $N = 2600$ subjects, representing the expected "average" image for a specific age. Known age-related effects, such as wider sulci and bigger ventricles (Fjell et al., 2009; Fjell and Walhovd, 2010), are clearly displayed.

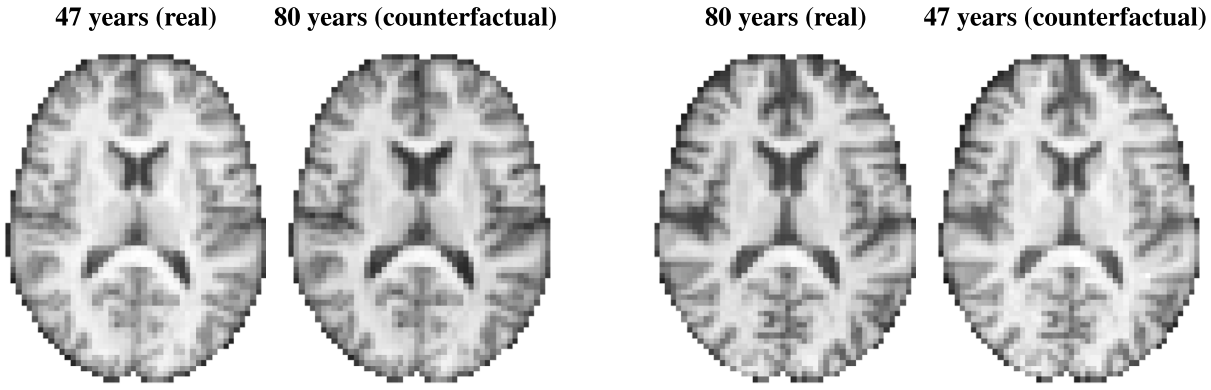


Fig. 14. Two examples of counterfactuals synthesized by the model for age prediction, trained on $N = 2600$ subjects. Left: real and counterfactual image of a 47-year old subject who is artificially aged to 80 years. Right: the same for a 80-year old subject who is rejuvenated to age 47.

test set, an internal 5-fold CV was performed within the training set, evaluating different values of K. The value with the highest classification accuracy was then retained and used to re-train the model on the entire training set, which was then evaluated on the test set. This procedure was repeated for each data split in the outer loop, to yield an overall prediction score.

We obtained an overall classification accuracy of 0.881. Fig. 15 shows the obtained spatial map $\mathbf{w}_G$ of the disease effect in one of the folds, together with the estimated age effect $\mathbf{w}_y$. Since the confounding effect of age is automatically controlled for in the model, $\mathbf{w}_G$ reflects the average brain changes that occur in direct response to AD disease (highlighting hippocampal atrophy in particular). For comparison, we repeated the same experiment but this time *without* including age as a confounder, simulating the setting of uncontrolled confounding shown in Fig. 1(d). Although the model still predicts well in this setting (classification accuracy 0.874), the estimated "disease effect" $\mathbf{w}_G$ is now misleading as it is biased to strongly reflect the effect of aging (see Fig. 16). The spatial map still indicates which image areas the method is paying attention to when making predictions, but now merely shows a statistical association rather than a direct causal relationship.

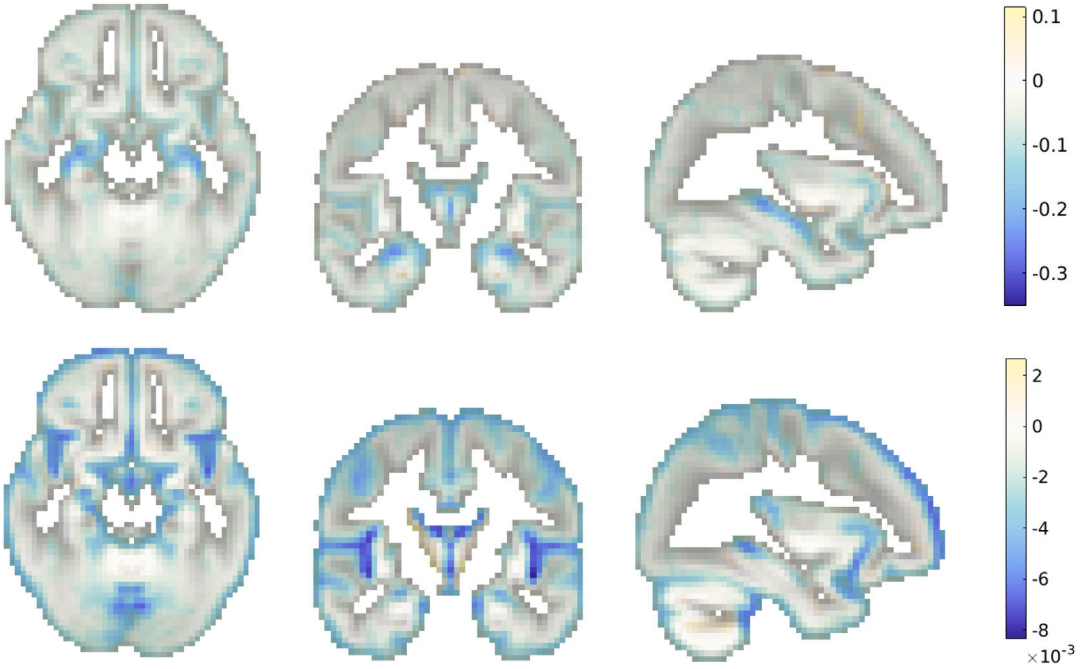


Fig. 15. Generative maps obtained in a disease classification experiment based on gray matter segmentations, when age is included as a confounder: $\mathbf{w}_G$ (top) and $\mathbf{w}_y$ (bottom) express the effect of AD and age, respectively. The maps are overlaid on the population template. Voxels with zero weight are transparent.

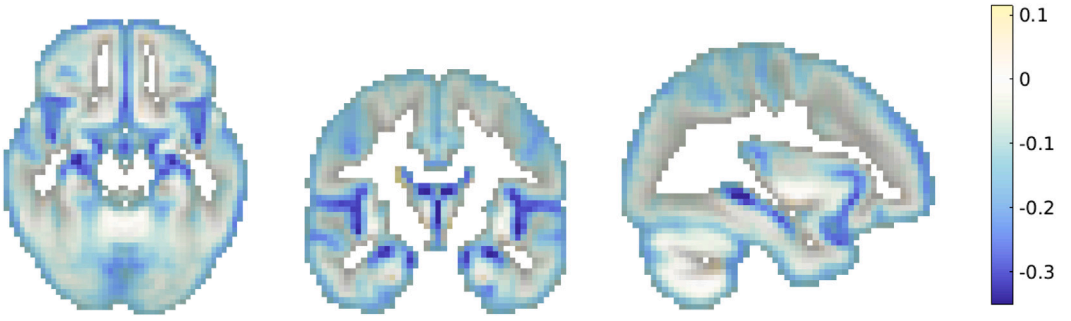


Fig. 16. Generative map $\mathbf{w}_G$ obtained in the same disease classification experiment shown of Fig. 15, but without controlling for age in the model. The map is now biased to also reflect the effect of aging instead of just the AD effect.

5.2. Nonlinearities in the causal model

Another extension of (1) is to consider nonlinear cause–effect relations in the model:

$$\mathbf{t} = \mathbf{m} + x\mathbf{w}_G + \sum_{l=1}^{L} f_l(x)\mathbf{w}_y^l + \boldsymbol{\eta},$$

where each $f_l(x)$ is some nonlinear function of the (continuous) variable of interest x. Although this corresponds to the simple causal diagram shown in Fig. 1(a), during training it can be viewed as a special case of the model extension of Section 5.1: the mappings $f_l(\cdot)$, $l = 1, \ldots, L$ can be evaluated for each training subject, and treated as known additional covariates in the model. Predicting with a trained model is no longer governed by the linear Eq. (6), though. We therefore invert the model by finely discretizing x^* into P distinct values $x_p, p = 1, \ldots, P$, and evaluating the posterior probability of each. Assuming a flat prior, this yields

$$p(x^* = x_p|\mathbf{t}^*, \mathbf{W}, \mathbf{C}) = \frac{\gamma_p}{\sum_{p'=1}^{P} \gamma_{p'}},$$

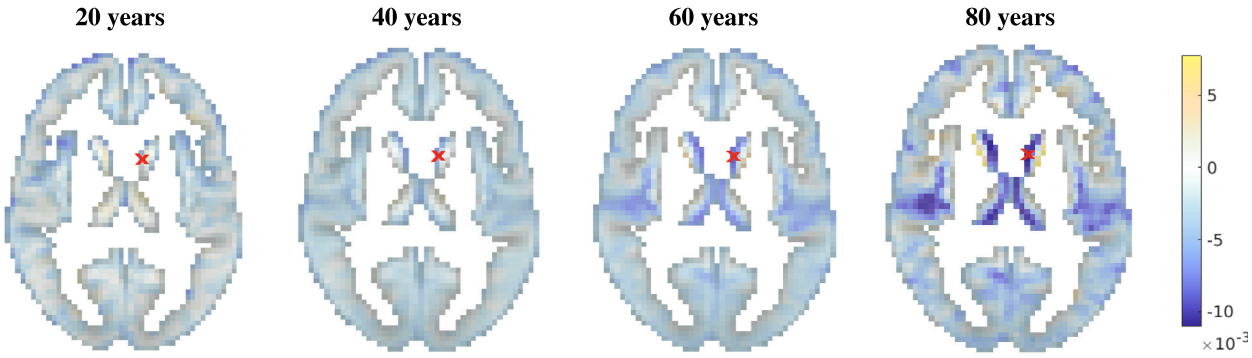
where $\gamma_p = \mathcal{N}(\mathbf{t}|\mathbf{m}+x_p\mathbf{w}_G+\sum_{l=1}^{L} f_l(x_p)\mathbf{w}_y^l, \mathbf{C})$ can be evaluated efficiently using (C.2), (C.3) and (C.4). Predictions are then obtained as the expected value of this posterior distribution:

$$\tilde{x}^* = \sum_{p=1}^{P} x_p p(x^* = x_p|\mathbf{t}^*, \mathbf{W}, \mathbf{C}).$$

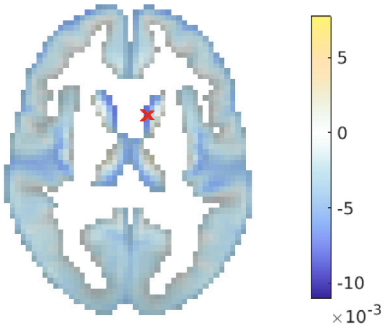
In order to demonstrate this variant, we tested whether the inclusion of an extra quadratic term in the forward model (i.e., $f_1(x) = x^2$) can improve age estimation results compared to the basic linear model. For this purpose, we used modulated gray matter segmentations (computed in the same way as described in Section 5.1) of T1-weighted scans of 562 healthy subjects from the IXI dataset.[7] This dataset was selected because aging has been shown to have an approximately quadratic effect across adulthood on some brain structures (Walhovd et al., 2005), and the IXI dataset covers an age span that is large enough (20–86 years) to possibly exploit this behavior for prediction purposes. In our experiment, we treated the possible inclusion of the quadratic term as an extra binary hyperparameter, determined in the same way as the other hyperparameter K of the method within a nested 5-fold CV (as in Section 5.1). For the discretization, we used $P = 20$ intervals covering the entire age range.

Using this set-up, we found that the quadratic model was selected in all the CV folds, and yielded significantly smaller age prediction test

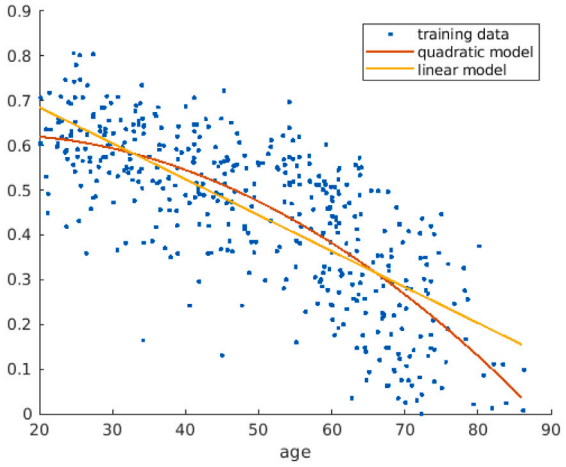
[7] https://brain-development.org/ixi-dataset/.



(a) Quadratic model



(b) Linear model



(c) Quadratic and linear fit in a selected voxel

Fig. 17. (a) Age-specific maps expressing the effect of aging, overlaid on the population template, when a quadratic aging model is used. These maps were obtained by computing for each voxel the tangent to the quadratic model at the specified age. Voxels with zero weight are transparent. (b) Corresponding generative map $\mathbf{w}_G$ expressing the age effect obtained when a basic linear model is used instead. (c) Training data in a selected voxel (marked in red in figures (a) and (b)), together with the fitted quadratic and linear models.

errors than the linear method (MAE of 4.36 vs. 4.73 years, p-value < 0.001 with a one-sided paired t-test). Fig. 17 shows linearized maps demonstrating the effect of aging, evaluated as several age points, for the quadratic model estimated in one of the folds (a), together with the corresponding map $\mathbf{w}_G$ obtained with the linear model (b). We observe that the aging effect is more pronounced in elderly subjects, as is also clearly displayed by the curve fitted in a single selected voxel shown in Fig. 17(c). In nonlinear cases such as this, subject-specific counterfactuals – such as the ones shown in Fig. 14 – remain straightforward to compute with the model, and may provide more useful individualized explanations than the group-level linearizations of Fig. 17(a).

6. Discussion

In this paper, we have proposed a lightweight method for image-based prediction that is inherently interpretable. It is based on classical human brain mapping techniques, but includes a multivariate noise model that yields accurate subject-level predictions when inverted. Despite its simplicity, the method predicts well in comparison with state-of-the art benchmarks, especially in typical training scenarios (those with no more than a few thousand training subjects) where more flexible techniques become prone to overfitting.

The method we described here represents a basic algorithm that can be further developed with more advanced techniques. For example, while we used external cross-validation to determine a suitable number of latent variables in the noise model, methods exist to infer this hyper-parameter automatically from the training data itself (Bishop, 1998). It should also be possible to address problems where the prediction targets x are not fully known in the training data, for instance by having the EM algorithm that currently estimates the noise covariance also infer missing mixture class memberships (Ghahramani and Hinton, 1996). Examples of such scenarios include disease classification tasks in which the "ground truth" diagnosis is noisy, or semi-supervised learning tasks in which a small training dataset is augmented with a large unlabeled one to improve prediction performance (Kingma et al., 2014). Finally, the generative model can be further extended to allow for more complex imaging data, such as a combination of multimodal images (Liem et al., 2017; Engemann et al., 2020; Cole, 2020), or longitudinal data where temporal correlations need to be explicitly taken into account (Bernal-Rusiel et al., 2013a,b).

This paper only addressed *encoding* scenarios in which an unknown condition x is assumed to be the cause of changes observed in imaging data $\mathbf{t}$ (see Fig. 1(a)–(c)). In this setting, it is well-known that only the forward model – represented by the generative maps $\mathbf{w}_G$ in the basic linear version of the proposed method – can be unambiguously interpreted, whereas methods working in the inverse, anti-causal direction (to predict x from $\mathbf{t}$) cannot: Their discriminative maps $\mathbf{w}_D$ may include areas merely to remove variations in $\mathbf{t}$ that are not due to x, while at the same time missing other areas affected by x that do not help in the prediction process (interpretation rules S7 and S4 in Weichwald et al. (2015), respectively). The contribution of this paper is therefore to provide users with access to a causal forward model, so that intuitive

explanations of the resulting prediction procedure can be generated. Unfortunately, this framework is not applicable in *decoding* scenarios where $\mathbf{t}$ causes x rather than vice versa (see Fig. 1(e)), excluding many important applications aimed at understanding how brain-based abnormalities give rise to symptoms. This is because the interpretation of such models works very differently: for instance, while $\mathbf{w}_D$ remains difficult to interpret directly (rules R3 and R4 in Weichwald et al. (2015)), areas highlighted in $\mathbf{w}_G$ no longer necessarily imply a causal relationship with x (rule R1).

Although we have used causal language to motivate a generative modeling approach to subject-level prediction, the generative maps depicted in this paper should not be over-interpreted as representing unbiased estimates of causality inferred from observational imaging data. Due to various measurement and selection biases, each model will necessarily be specific only to the dataset it was derived from, without generalizing to other datasets the way a population-level unbiased model would (Wachinger et al., 2021). For instance, the UK Biobank cohort is known to not be representative of the general population with regard to a number of sociodemographic, physical, lifestyle, and health-related characteristics (Fry et al., 2017), and this will be reflected in the generative maps that are produced. Similarly, there is a strong dependency of MRI-derived features on such factors as the scanners, the pulse sequences and the image processing pipelines that are used, which will also find their way into the model estimates. Finally, a key assumption to preserve interpretability in the proposed model is that all confounders are known, so that the situation depicted in Fig. 1(d) can be excluded. Although the requirement of affecting both a patient's condition and their imaging data independently dramatically reduces the number of potential confounders (for instance, of the hundreds of potential confounds studied for UK Biobank brain imaging (Alfaro-Almagro et al., 2021), very few beyond age, gender and perhaps imaging site would likely qualify), we may not know all relevant factors nor have data on them in practice, resulting in confounding bias (Wachinger et al., 2021).

Although this paper focused on interpretable *models* for prediction, it is worth remembering that the interpretability of the resulting system also depends on the type of *features* that are used. In the examples given throughout this paper, the features consisted of voxel-level intensities obtained after warping each subject into a common template space with nonlinear registration. While nonlinear registration helps remove irrelevant inter-subject shape variations from the data, it can also hide biological effects: In the age-conditioned synthesized images of Figs. 13 and 14, for instance, a significant portion of the atrophy patterns caused by aging (namely the part encoded in the nonlinear warps) is missing. This could potentially be avoided by using the parameters governing the warps – such as the stationary velocity fields in diffeomorphic registration models (Ashburner, 2007) – as (additional) features in the proposed model.

A related issue is how much of a prediction method's robustness to inter-subject variability should be the responsibility of data preprocessing vs. that of the prediction model itself. It could be argued that, in our experiments, much of the "heavy lifting" was done by the preprocessing so that even simple linear-Gaussian models could work well. This point, however, is nuanced by the fact that some form of preprocessing is currently required for all subject-level prediction methods in the literature. In the SFCN paper, for instance, it is demonstrated that a neural network can "model away" nonlinear deformations that have not been removed from the input images; however the images have still been preprocessed with affine registration and a full segmentation pipeline (paradoxally involving nonlinear registration) for skull stripping, bias field correction and intensity normalization.

The question of data preprocessing vs. modeling is further complicated by the issue of scanner- and sequence-dependent MRI contrast in real-world applications: A subject-level prediction method that is trained on raw intensities will rarely be directly applicable to data acquired elsewhere. In principle, all these issues can be addressed by integrating the type of forward models used in Bayesian segmentation (Ashburner and Friston, 2005; Puonti et al., 2016) – which include subject-specific template deformation, bias field correction and contrast-adaptive intensity modeling — within the proposed generative approach itself. Whether this will yield a tangible benefit compared to preprocessing the data depends on the difficulty of training and inverting such integrated models in practice. As demonstrated by the VAE example analyzed in this paper, if building higher-capacity models requires making approximations to keep computations feasible, prediction performance may be hurt rather than helped.

In conclusion, we have proposed a lightweight generative model that is inherently interpretable and that can still make accurate predictions. Since it can easily be extended, we hope it will form a useful basis for future developments in the field.

CRediT authorship contribution statement

Chiara Mauri: Writing – review & editing, Writing - original draft, Visualization, Validation, Software, Resources, Methodology, Investigation, Data curation, Conceptualization. **Stefano Cerri:** Writing – review & editing, Validation, Software. **Oula Puonti:** Writing – review & editing, Methodology, Conceptualization. **Mark Mühlau:** Writing – review & editing, Supervision. **Koen Van Leemput:** Writing – review & editing, Supervision, Project administration, Methodology, Conceptualization.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Oula Puonti reports financial support was provided by Lundbeck Foundation. Koen Van Leemput is regular member of the editorial board of Medical Image Analysis

Acknowledgments

This research was conducted using the UK Biobank Resource under Application Number 65657, and it was made possible in part by the computational hardware generously provided by the Massachusetts Life Sciences Center (https://www.masslifesciences.com/). This work was supported by the National Institute of Neurological Disorders and Stroke, USA under grant number R01NS112161, and by the CLAIMS project, which is supported by the Innovative Health Initiative Joint Undertaking (JU) under grant agreement No 101112153. The JU receives support from the European Union's Horizon Europe research and innovation programme and COCIR, EFPIA, EuropaBio, MedTech Europe, Vaccines Europe, AB Science SA and Icometrix NV. OP is supported by the Lundbeck Foundation, Denmark (grant R360-2021-395).

Appendix A. Derivations of model inversion

Here we derive the expressions for making predictions about the variable of interest. For a binary target variable x^*,

$$
\begin{aligned}
&p(x^*=1|\mathbf{t}^*,\mathbf{W},\mathbf{C}) \\
&\quad= \frac{p(\mathbf{t}^*|x^*=1,\mathbf{W},\mathbf{C})p(x^*=1)}{p(\mathbf{t}^*|x^*=1,\mathbf{W},\mathbf{C})p(x^*=1)+p(\mathbf{t}^*|x^*=0,\mathbf{W},\mathbf{C})p(x^*=0)} \\
&\quad= \frac{1}{1+\frac{p(t^*|x^*=0,\boldsymbol{W},\boldsymbol{C})p(x^*=0)}{p(t^*|x^*=1,\boldsymbol{W},\boldsymbol{C})p(x^*=1)}} \\
&\quad= \sigma\Big[\log p(\mathbf{t}^*|x^*=1,\mathbf{W},\mathbf{C}) - \log p(\mathbf{t}^*|x^*=0,\mathbf{W},\mathbf{C}) \\
&\qquad + \log p(x^*=1) - \log p(x^*=0)\Big],
\end{aligned}
$$

where

$$\log p(\mathbf{t}^*|x^*=1,\mathbf{W},\mathbf{C}) - \log p(\mathbf{t}^*|x^*=0,\mathbf{W},\mathbf{C})$$

$$= -\frac{1}{2}(\mathbf{t}^* - \mathbf{m} - \mathbf{w}_G)^T \mathbf{C}^{-1}(\mathbf{t}^* - \mathbf{m} - \mathbf{w}_G) + \frac{1}{2}(\mathbf{t}^* - \mathbf{m})^T \mathbf{C}^{-1}(\mathbf{t}^* - \mathbf{m}) = \mathbf{w}_G^T \mathbf{C}^{-1}(\mathbf{t}^* - \mathbf{m}) - \frac{1}{2}\mathbf{w}_G^T \mathbf{C}^{-1}\mathbf{w}_G.$$

This explains (2).

For a continuous target variable with flat prior, the log-posterior

$$\log p(x^*|\mathbf{t}^*, \mathbf{W}, \mathbf{C}) = -\frac{1}{2}(\mathbf{t}^* - \mathbf{m} - x^*\mathbf{w}_G)^T \mathbf{C}^{-1}(\mathbf{t}^* - \mathbf{m} - x^*\mathbf{w}_G) + \text{const}$$

is quadratic with derivative

$$\frac{d \log p(x^*|\mathbf{t}^*, \mathbf{W}, \mathbf{C})}{dx^*} = \mathbf{w}_G^T \mathbf{C}^{-1}(\mathbf{t}^* - \mathbf{m} - x^*\mathbf{w}_G) \tag{A.1}$$

and curvature

$$\frac{d^2 \log p(x^*|\mathbf{t}^*, \mathbf{W}, \mathbf{C})}{dx^{*2}} = -\mathbf{w}_G^T \mathbf{C}^{-1}\mathbf{w}_G.$$

Therefore, the posterior is Gaussian, with variance given by (5). The mean is obtained by setting (A.1) to zero, which yields (6).

Appendix B. Expression for W

For training, the log marginal likelihood is given by

$$\log p\left(\{\mathbf{t}_n\}_{n=1}^N | \{x_n\}_{n=1}^N, \mathbf{W}, \mathbf{C}\right) = \sum_{n=1}^N -\frac{1}{2}(\mathbf{t}_n - \mathbf{W}\boldsymbol{\phi}_n)^T \mathbf{C}^{-1}(\mathbf{t}_n - \mathbf{W}\boldsymbol{\phi}_n) + \text{const},$$

which has as gradient with respect to $\mathbf{W}$:

$$\sum_{n=1}^N \mathbf{C}^{-1}(\mathbf{t}_n - \mathbf{W}\boldsymbol{\phi}_n)\boldsymbol{\phi}_n^T.$$

Setting to zero and re-arranging yields (8).

Appendix C. Efficient implementation

Using Woodbury's identity, we obtain

$$\mathbf{C}^{-1} = \boldsymbol{\Delta}^{-1} - \boldsymbol{\Delta}^{-1}\mathbf{V}\left(\mathbb{I}_K + \mathbf{V}^T\boldsymbol{\Delta}^{-1}\mathbf{V}\right)^{-1}\mathbf{V}^T\boldsymbol{\Delta}^{-1} = \boldsymbol{\Delta}^{-1} - \boldsymbol{\Delta}^{-1}\mathbf{V}\boldsymbol{\Sigma}\mathbf{V}^T\boldsymbol{\Delta}^{-1}, \tag{C.1}$$

and therefore (3) can be computed as

$$\mathbf{w}_D = \boldsymbol{\Delta}^{-1}\mathbf{w}_G - \boldsymbol{\Delta}^{-1}\mathbf{V}\boldsymbol{\Sigma}\left(\mathbf{V}^T\boldsymbol{\Delta}^{-1}\mathbf{w}_G\right).$$

Using this result, (5) is given by $v = 1/(\mathbf{w}_D^T\mathbf{w}_G)$.

Computing the marginal likelihood (7) – which is needed both to monitor convergence of the EM algorithm during model training, and to invert the model with nonlinearities (Section 5.2) – involves numerical evaluations of the form:

$$\log \mathcal{N}(\boldsymbol{\eta}|\, \mathbf{0}, \mathbf{C}) \propto \boldsymbol{\eta}^T\mathbf{C}^{-1}\boldsymbol{\eta} + \log|\mathbf{C}| + \text{const}. \tag{C.2}$$

Using (C.1), the first term can be computed as

$$\boldsymbol{\eta}^T\mathbf{C}^{-1}\boldsymbol{\eta} = \boldsymbol{\eta}^T\boldsymbol{\Delta}^{-1}(\boldsymbol{\eta} - \mathbf{V}\boldsymbol{\Sigma}\mathbf{V}^T\boldsymbol{\Delta}^{-1}\boldsymbol{\eta}) = \boldsymbol{\eta}^T\boldsymbol{\Delta}^{-1}(\boldsymbol{\eta} - \mathbf{V}\boldsymbol{\mu}), \tag{C.3}$$

with $\boldsymbol{\mu} = \boldsymbol{\Sigma}\mathbf{V}^T\boldsymbol{\Delta}^{-1}\boldsymbol{\eta}$ being an estimate of the latent variables. The second term can be computed using Sylvester's determinant identity (Akritas et al., 1996):

$$|\mathbf{V}\mathbf{V}^T\boldsymbol{\Delta}^{-1} + \mathbb{I}_J| = |\mathbf{V}^T\boldsymbol{\Delta}^{-1}\mathbf{V} + \mathbb{I}_K| = |\boldsymbol{\Sigma}|^{-1},$$

so that

$$\log|\mathbf{C}| = \log|\boldsymbol{\Delta}| - \log|\boldsymbol{\Sigma}|. \tag{C.4}$$

Finally, the EM update (13) of the diagonal matrix $\boldsymbol{\Delta}$ can be computed one element at a time: the diagonal element corresponding to the jth voxel is given by

$$\begin{aligned}\boldsymbol{\Delta}_{jj} &= \frac{1}{N}\mathbf{r}_j^T\left(\mathbf{r}_j - \mathbf{M}\mathbf{v}_j\right)\\ &= \frac{1}{N}\|\mathbf{r}_j - \mathbf{M}\mathbf{v}_j\|^2 + \frac{1}{N}\mathbf{r}_j^T\mathbf{M}\mathbf{v}_j - \frac{1}{N}\mathbf{v}_j^T\mathbf{M}^T\mathbf{M}\mathbf{v}_j\\ &= \frac{1}{N}\|\mathbf{r}_j - \mathbf{M}\mathbf{v}_j\|^2 + \frac{1}{N}\mathbf{v}_j^T\left(\mathbf{M}^T\mathbf{M} + N\boldsymbol{\Sigma}\right)\mathbf{v}_j - \frac{1}{N}\mathbf{v}_j^T\mathbf{M}^T\mathbf{M}\mathbf{v}_j\\ &= \frac{1}{N}\|\mathbf{r}_j - \mathbf{M}\mathbf{v}_j\|^2 + \mathbf{v}_j^T\boldsymbol{\Sigma}\mathbf{v}_j.\end{aligned}$$

Here $\mathbf{v}_j^T$ is the jth row of $\mathbf{V}$, $\mathbf{M} = \left(\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_N\right)^T$, and $\mathbf{r}_j = (\eta_1{}^j, \ldots, \eta_N{}^j)^T$ where $\eta_n{}^j$ denotes the jth element of $\boldsymbol{\eta}_n$. The next to last step makes use of the fact that (see (12))

$$\mathbf{v}_j = \left(\mathbf{M}^T\mathbf{M} + N\boldsymbol{\Sigma}\right)^{-1}\mathbf{M}^T\mathbf{r}_j.$$

Appendix D. Analysis of the Haufe transformation

When the target variable x has zero mean in the training data, i.e., $(\sum_{n=1}^N x_n)/N = 0$, the solution of (8) is given by

$$\mathbf{m} = \frac{\sum_{n=1}^N \mathbf{t}_n}{N}$$

and

$$\mathbf{w}_G = \left(\sum_{n=1}^N \mathbf{t}_n x_n\right)\left(\sum_{n=1}^N x_n^2\right)^{-1}, \tag{D.1}$$

because

$$\sum_{n=1}^N \boldsymbol{\phi}_n\boldsymbol{\phi}_n^T = \begin{pmatrix} N & 0 \\ 0 & \sum_{n=1}^N x_n^2 \end{pmatrix}.$$

Using the notation $\mathbf{T} = \left(\mathbf{t}_1, \ldots, \mathbf{t}_N\right)$ and $\mathbf{x} = (x_1, \ldots, x_n)^T$, (D.1) can also be written as

$$\mathbf{w}_G = \mathbf{T}\mathbf{x}(\mathbf{x}^T\mathbf{x})^{-1}. \tag{D.2}$$

Given some linear discriminative method with weight vector $\mathbf{w}$, the method in Haufe et al. (2014) aims to recover a corresponding generative weight vector $\tilde{\mathbf{w}}_G$ as follows. Rather than using the available *real* training targets $\{x_n\}$ – which were used to obtain $\mathbf{w}$ – (D.2) is instead applied to *estimated* targets $\{\tilde{x}_n\}$ obtained with the discriminative method:

$$\tilde{x}_n = \mathbf{w}^T\mathbf{t}_n, \quad \forall n,$$

where it is assumed that the training images have been preprocessed to have zero mean ($\mathbf{m} = \mathbf{0}$). Writing $\tilde{\mathbf{x}} = \mathbf{T}^T\mathbf{w}$ and plugging this into (D.2) then yields

$$\tilde{\mathbf{w}}_G = \mathbf{T}\tilde{\mathbf{x}}(\tilde{\mathbf{x}}^T\tilde{\mathbf{x}})^{-1} = \mathbf{T}(\mathbf{T}^T\mathbf{w})(\tilde{\mathbf{x}}^T\tilde{\mathbf{x}})^{-1} = \boldsymbol{\Sigma}_t\mathbf{w}\tilde{\sigma}_x^{-2}$$

where $\boldsymbol{\Sigma}_t = N^{-1}\mathbf{T}\mathbf{T}^T$ and $\tilde{\sigma}_x^2 = N^{-1}\tilde{\mathbf{x}}^T\tilde{\mathbf{x}}$. This corresponds to (6) in Haufe et al. (2014).

A fundamental issue with the obtained expression

$$\tilde{\mathbf{w}}_G = \boldsymbol{\Sigma}_t\mathbf{w}\tilde{\sigma}_x^{-2}$$

(the Haufe transformation) is that it obscures the fact that the dependency of $\tilde{\mathbf{w}}_G$ on $\mathbf{w}$ is only through the predictions $\tilde{\mathbf{x}} = \mathbf{T}^T\mathbf{w}$. This is a problem, because good predictions can be obtained with very different $\mathbf{w}$'s – including those ignoring entire image areas – which prevents $\tilde{\mathbf{w}}_G$ from being a reliable indication of feature importance. To illustrate this point, in Fig. D.18 we trained two different $\mathbf{w}$'s to predict age in the UK Biobank dataset used in Section 4 with the RVoxM method, while clamping the weights w_j to zero in large (and very different) image areas that are therefore necessarily ignored when

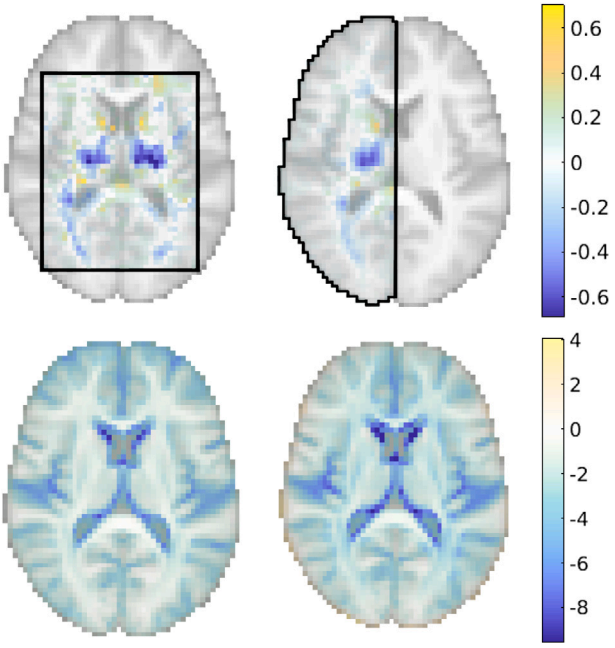


Fig. D.18. Illustration of why the Haufe transformation should not be used to explain the prediction process of a discriminative method. Top: weights obtained when RVoxM is trained to predict age from $N = 2600$ subjects when the weights are clamped to zero outside of two different ROIs indicated by the black outline. On the left the ROI encompasses only a cuboid area around the ventricles, whereas on the right only one hemisphere is included. Bottom: the corresponding maps $\tilde{\mathbf{w}}_G$ obtained with Haufe's transformation hide the fact that voxels outside of the respective ROIs are never used in the predictions.

making predictions. Since both regimes still predict well (MAE within 6% of what is obtained with RVoxM without clamping), the resulting $\tilde{\mathbf{w}}_G$'s look very similar to each other and to the corresponding map $\mathbf{w}_G$ shown in Fig. 11(middle), completely failing to reveal that certain image areas are never actually looked at.

Appendix E. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.media.2024.103436.

Data availability

The code is available online. The UK Biobank dataset is accessible upon applications via the website: https://www.ukbiobank.ac.uk/.

References

Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., Kim, B., 2018. Sanity checks for saliency maps. Adv. Neural Inf. Process. Syst. 31.

Akritas, A., Akritas, E., Malaschonok, G., 1996. Various proofs of Sylvester's (determinant) identity. Math. Comput. Simulation 42, 585–593.

Alfaro-Almagro, F., Jenkinson, M., Bangerter, N.K., Andersson, J.L., Griffanti, L., Douaud, G., Sotiropoulos, S.N., Jbabdi, S., Hernandez-Fernandez, M., Vallee, E., Vidaurre, D., Webster, M., McCarthy, P., Rorden, C., Daducci, A., Alexander, D.C., Zhang, H., Dragonu, I., Matthews, P.M., Miller, K.L., Smith, S.M., 2018. Image processing and quality control for the first 10,000 brain imaging datasets from UK Biobank. Neuroimage 166, 400–424.

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

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

Arrieta, A.B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., Chatila, R., Herrera, F., 2020. Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Inf. Fusion 58, 82–115.

Arun, N., Gaw, N., Singh, P., Chang, K., Aggarwal, M., Chen, B., Hoebel, K., Gupta, S., Patel, J., Gidwani, M., Adebayo, J., Li, M.D., Kalpathy-Cramer, J., 2021. Assessing the trustworthiness of saliency maps for localizing abnormalities in medical imaging. Radiol.: Artif. Intell. 3 (6), e200267.

Ashburner, J., 2007. A fast diffeomorphic image registration algorithm. Neuroimage 38 (1), 95–113.

Ashburner, J., Barnes, G., Chen, C.-C., Daunizeau, J., Flandin, G., Friston, K., Gitelman, D., Glauche, V., Henson, R., Hutton, C., Jafarian, A., Kiebel, S., Kilner, J., Litvak, V., Moran, R., Penny, W., Phillips, C., Razi, A., Stephan, K., Tak, S., Tyrer, A., Zeidman, P., 2014. SPM12 manual. Wellcome Trust Cent. Neuroimaging Lond. UK 2464 (4).

Ashburner, J., Friston, K.J., 2000. Voxel-based morphometry–the methods. Neuroimage 11 (6), 805–821.

Ashburner, J., Friston, K.J., 2005. Unified segmentation. Neuroimage 26 (3), 839–851.

Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.-R., Samek, W., 2015. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PLoS One 10 (7), e0130140.

Baehrens, D., Schroeter, T., Harmeling, S., Kawanabe, M., Hansen, K., Müller, K.-R., 2010. How to explain individual classification decisions. J. Mach. Learn. Res. 11, 1803–1831.

Bernal-Rusiel, J.L., Greve, D.N., Reuter, M., Fischl, B., Sabuncu, M.R., 2013a. Statistical analysis of longitudinal neuroimage data with linear mixed effects models. Neuroimage 66, 249–260.

Bernal-Rusiel, J.L., Reuter, M., Greve, D.N., Fischl, B., Sabuncu, M.R., 2013b. Spatiotemporal linear mixed effects modeling for the mass-univariate analysis of longitudinal neuroimage data. Neuroimage 81, 358–370.

Bishop, C., 1998. Bayesian pca. Adv. Neural Inf. Process. Syst. 11.

Bishop, C.M., 2006. Pattern Recognition and Machine Learning. Springer.

Breteler, M.M., Stöcker, T., Pracht, E., Brenner, D., Stirnberg, R., 2014. IC-P-165: MRI in the rhineland study: A novel protocol for population neuroimaging. Alzheimer's Dement. 10, 92.

Chevalier, J.-A., Nguyen, T.-B., Salmon, J., Varoquaux, G., Thirion, B., 2021. Decoding with confidence: Statistical control on decoder maps. Neuroimage 234, 117921.

Chung, M.K., Worsley, K.J., Paus, T., Cherif, C., Collins, D.L., Giedd, J.N., Rapoport, J.L., Evans, A.C., 2001. A unified statistical approach to deformation-based morphometry. Neuroimage 14 (3), 595–606.

Cole, J.H., 2020. Multimodality neuroimaging brain-age in UK biobank: relationship to biomedical, lifestyle, and cognitive factors. Neurobiol. Aging 92, 34–42.

Cole, J.H., Franke, K., Cherbuin, N., 2019. Quantification of the biological age of the brain using neuroimaging. In: Biomarkers of Human Aging. Springer, pp. 293–328.

Dalca, A., Rakic, M., Guttag, J., Sabuncu, M., 2019. Learning conditional deformable templates with convolutional networks. Adv. Neural Inf. Process. Syst. 32.

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

Di Martino, A., Yan, C.-G., Li, Q., Denio, E., Castellanos, F.X., Alaerts, K., Anderson, J.S., Assaf, M., Bookheimer, S.Y., Dapretto, M., Deen, B., Delmonte, S., Dinstein, I., Ertl-Wagner, B., Fair, D.A., Gallagher, L., Kennedy, D.P., Keown, C.L., Keysers, C., Lainhart, J.E., Lord, C., Luna, B., Menon, V., Minshew, N.J., Monk, C.S., Mueller, S., Müller, R.-A., Nebel, M.B., Nigg, J.T., O'Hearn, K., Pelphrey, K.A., Peltier, S.J., Rudie, J.D., Sunaert, S., Thioux, M., Tyszka, J.M., Uddin, L.Q., Verhoeven, J.S., Wenderoth, N., Wiggins, J.L., Mostofsky, S.H., Milham, M.P., 2014. The autism brain imaging data exchange: Towards a large-scale evaluation of the intrinsic brain architecture in autism. Mol. Psychiatry 19 (6), 659–667.

Domingos, P., 2012. A few useful things to know about machine learning. Commun. ACM 55 (10), 78–87.

Domingos, P., Pazzani, M., 1997. On the optimality of the simple Bayesian classifier under zero-one loss. Mach. Learn. 29 (2), 103–130.

Ellis, K.A., Bush, A.I., Darby, D., De Fazio, D., Foster, J., Hudson, P., Lautenschlager, N.T., Lenzo, N., Martins, R.N., Maruff, P., Masters, C., Milner, A., Pike, K., Rowe, C., Savage, G., Szoeke, C., Taddei, K., Villemagne, V., Woodward, M., Ames, D., 2009. The Australian Imaging, Biomarkers and Lifestyle (AIBL) study of aging: Methodology and baseline characteristics of 1112 individuals recruited for a longitudinal study of Alzheimer's disease. Int. Psychogeriatr. 21 (4), 672–687.

Engemann, D.A., Kozynets, O., Sabbagh, D., Lemaître, G., Varoquaux, G., Liem, F., Gramfort, A., 2020. Combining magnetoencephalography with magnetic resonance imaging enhances learning of surrogate-biomarkers. Elife 9, e54055.

Fischl, B., Dale, A.M., 2000. Measuring the thickness of the human cerebral cortex from magnetic resonance images. Proc. Natl. Acad. Sci. USA 97 (20), 11050–11055.

Fjell, A.M., Walhovd, K.B., 2010. Structural brain changes in aging: Courses, causes and cognitive consequences. Rev. Neurosci. 21 (3), 187–222.

Fjell, A.M., Westlye, L.T., Amlien, I., Espeseth, T., Reinvang, I., Raz, N., Agartz, I., Salat, D.H., Greve, D.N., Fischl, B., Dale, A.M., Walhovd, K.B., 2009. High consistency of regional cortical thinning in aging across multiple samples. Cerebral Cortex 19 (9), 2001–2012. http://dx.doi.org/10.1093/cercor/bhn232.

Friston, K., Chu, C., Mourão-Miranda, J., Hulme, O., Rees, G., Penny, W., Ashburner, J., 2008. Bayesian decoding of brain images. Neuroimage 39 (1), 181–205.
Friston, K.J., Frith, C., Liddle, P., Frackowiak, R., 1991. Comparing functional (PET) images: the assessment of significant change. J. Cereb. Blood Flow Metab. 11 (4), 690–699.
Friston, K.J., Holmes, A.P., Worsley, K.J., Poline, J.-P., Frith, C.D., Frackowiak, R.S., 1994. Statistical parametric maps in functional imaging: A general linear approach. Hum. Brain Mapp. 2 (4), 189–210.
Fry, A., Littlejohns, T.J., Sudlow, C., Doherty, N., Adamska, L., Sprosen, T., Collins, R., Allen, N.E., 2017. Comparison of sociodemographic and health-related characteristics of UK biobank participants with those of the general population. Am. J. Epidemiol. 186 (9), 1026–1034.
German National Cohort Consortium, 2014. The german national cohort: Aims, study design and organization. Eur. J. Epidemiol. 29 (5), 371–382.
Ghahramani, Z., Hinton, G.E., 1996. The EM Algorithm for Mixtures of Factor Analyzers. Technical report, Citeseer.
Ghassemi, M., Oakden-Rayner, L., Beam, A.L., 2021. The false hope of current approaches to explainable artificial intelligence in health care. Lancet Digit. Health 3 (11), e745–e750.
Gu, J., Tresp, V., 2019. Saliency methods for explaining adversarial attacks. arXiv preprint arXiv:1908.08413.
Hart, P.E., Stork, D.G., Duda, R.O., 2000. Pattern Classification. Wiley Hoboken.
Haufe, S., Meinecke, F., Görgen, K., Dähne, S., Haynes, J.-D., Blankertz, B., Bießmann, F., 2014. On the interpretation of weight vectors of linear models in multivariate neuroimaging. Neuroimage 87, 96–110.
Jack, Jr., C.R., Bernstein, M.A., Fox, N.C., Thompson, P., Alexander, G., Harvey, D., Borowski, B., Britson, P.J., L. Whitwell, J., Ward, C., Dale, A.M., Felmlee, J.P., Gunter, J.L., Hill, D.L., Killiany, R., Schuff, N., Fox-Bosetti, S., Lin, C., Studholme, C., DeCarli, C.S., Krueger, G., Ward, H.A., Metzger, G.J., Scott, K.T., Mallozzi, R., Blezek, D., Levy, J., Debbins, J.P., Fleisher, A.S., Albert, M., Green, R., Bartzokis, G., Glover, G., Mugler, J., Weiner, M.W., 2008. The Alzheimer's Disease neuroimaging initiative (ADNI): MRI methods. J. Magn. Reson. Imaging: Off. J. Int. Soc. Magn. Reson. Med. 27 (4), 685–691.
Joint Epilepsy Council, 2011. Epilepsy Prevalence, Incidence and Other Statistics. Joint Epilepsy Council, UK and Ireland, URL https://d3imrogdy81qei.cloudfront.net/instructor_docs/373/29_05_2016_Joint_Epilepsy_Council_Prevalence_and_Incidence_September_11.pdf.
Kaufmann, T., van der Meer, D., Doan, N.T., Schwarz, E., Lund, M.J., Agartz, I., Alnæs, D., Barch, D.M., Baur-Streubel, R., Bertolino, A., Bettella, F., Beyer, M.K., Bøen, E., Borgwardt, S., Brandt, C.L., Buitelaar, J., Celius, E.G., Cervenka, S., Conzelmann, A., Córdova-Palomera, A., Dale, A.M., de Quervain, D.J.F., Di Carlo, P., Djurovic, S., Dørum, E.S., Eisenacher, S., Elvsåshagen, T., Espeseth, T., Fatouros-Bergman, H., Flyckt, L., Franke, B., Frei, O., Haatveit, B., Håberg, A.K., Harbo, H.F., Hartman, C.A., Heslenfeld, D., Hoekstra, P.J., Høgestøl, E.A., Jernigan, T.L., Jonassen, R., Jönsson, E.G., Karolinska Schizophrenia Project (KaSP), P., Kłoszewska, I., Kolskår, K.K., Landrø, N.I., Le Hellard, S., Lesch, K.-P., Lovestone, S., Lundervold, A., Lundervold, A.J., Maglanoc, L.A., Malt, U.F., Mecocci, P., Melle, I., Meyer-Lindenberg, A., Moberget, T., Norbom, L.B., Nordvik, J.E., Nyberg, L., Oosterlaan, J., Papalino, M., Papassotiropoulos, A., Pauli, P., Pergola, G., Persson, K., Richard, G., Rokicki, J., Sanders, A.-M., Selbæk, G., Shadrin, A.A., Smeland, H., Sowa, P., Steen, V.M., Tsolaki, M., Ulrichsen, K.M., Vellas, B., Wang, L., Westman, E., Ziegler, G.C., Zink, M., Andreassen, O.A., Westlye, L.T., 2019. Common brain disorders are associated with heritable patterns of apparent aging of the brain. Nat. Neurosci. 22 (10), 1617–1623.
Kingma, D.P., Mohamed, S., Jimenez Rezende, D., Welling, M., 2014. Semi-supervised learning with deep generative models. Adv. Neural Inf. Process. Syst. 27.
Kingma, D.P., Welling, M., 2013. Auto-encoding variational Bayes. arXiv preprint arXiv:1312.6114.
Liem, F., Varoquaux, G., Kynast, J., Beyer, F., Masouleh, S.K., Huntenburg, J.M., Lampe, L., Rahim, M., Abraham, A., Craddock, R.C., et al., 2017. Predicting brain-age from multimodal imaging data captures cognitive impairment. Neuroimage 148, 179–188.
Lipton, Z.C., 2018. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. Queue 16 (3), 31–57.
Mackenzie, I., Morant, S., Bloomfield, G., MacDonald, T., O'riordan, J., 2014. Incidence and prevalence of multiple sclerosis in the UK 1990–2010: A descriptive study in the General Practice Research Database. J. Neurol. Neurosurg. Psychiatry 85 (1), 76–84.
Mauri, C., Cerri, S., Puonti, O., Mühlau, M., Van Leemput, K., 2022. Accurate and explainable image-based prediction using a lightweight generative model. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 448–458.
Miller, K.L., Alfaro-Almagro, F., Bangerter, N.K., Thomas, D.L., Yacoub, E., Xu, J., Bartsch, A.J., Jbabdi, S., Sotiropoulos, S.N., Andersson, J.L., Griffanti, L., Douaud, G., Okell, T.W., Weale, P., Dragonu, I., Garratt, S., Hudson, S., Collins, R., Jenkinson, M., Matthews, P.M., Smith, S.M., 2016. Multimodal population brain imaging in the UK Biobank prospective epidemiological study. Nat. Neurosci. 19 (11), 1523–1536.
Mouches, P., Wilms, M., Rajashekar, D., Langner, S., Forkert, N.D., 2022. Multimodal biological brain age prediction using Magnetic Resonance Imaging and angiography with the identification of predictive regions. Hum. Brain Mapp. 43 (8), 2554–2566.
Ng, A., Jordan, M., 2002. On discriminative vs. generative classifiers: A comparison of logistic regression and naive Bayes. In: Advances in Neural Information Processing Systems. pp. 841–848.
Pawlowski, N., Coelho de Castro, D., Glocker, B., 2020. Deep structural causal models for tractable counterfactual inference. Adv. Neural Inf. Process. Syst. 33, 857–869.
Pearl, J., Mackenzie, D., 2018. The Book of Why: The New Science of Cause and Effect. Basic books.
Peng, H., Gong, W., Beckmann, C.F., Vedaldi, A., Smith, S.M., 2021. Accurate brain age prediction with lightweight deep neural networks. Med. Image Anal. 68, 101871.
Pinaya, W.H., Tudosiu, P.-D., Dafflon, J., Da Costa, P.F., Fernandez, V., Nachev, P., Ourselin, S., Cardoso, M.J., 2022. Brain imaging generation with latent diffusion models. In: MICCAI Workshop on Deep Generative Models. Springer, pp. 117–126.
Pölsterl, S., Wachinger, C., 2021. Estimation of causal effects in the presence of unobserved confounding in the Alzheimer's continuum. In: Information Processing in Medical Imaging: 27th International Conference, IPMI 2021, Virtual Event, June 28–June 30, 2021, Proceedings 27. Springer, pp. 45–57.
Puonti, O., Iglesias, J.E., Van Leemput, K., 2016. Fast and sequence-adaptive whole-brain segmentation using parametric Bayesian modeling. Neuroimage 143, 235–249.
Ras, G., Xie, N., Van Gerven, M., Doran, D., 2022. Explainable deep learning: A field guide for the uninitiated. J. Artificial Intelligence Res. 73, 329–397.
Ravi, D., Alexander, D.C., Oxtoby, N.P., Alzheimer's Disease Neuroimaging Initiative, 2019. Degenerative adversarial neuroimage nets: Generating images that mimic disease progression. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 164–172.
Rubin, D.B., Thayer, D.T., 1982. EM algorithms for ML factor analysis. Psychometrika 47 (1), 69–76.
Rudin, C., 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat. Mach. Intell. 1 (5), 206–215.
Sabuncu, M.R., Van Leemput, K., 2012. The relevance voxel machine (RVoxM): A self-tuning Bayesian model for informative image-based prediction. IEEE Trans. Med. Imaging 31 (12), 2290–2306.
Satterthwaite, T.D., Elliott, M.A., Ruparel, K., Loughead, J., Prabhakaran, K., Calkins, M.E., Hopson, R., Jackson, C., Keefe, J., Riley, M., et al., 2014. Neuroimaging of the Philadelphia neurodevelopmental cohort. Neuroimage 86, 544–553.
Schram, M.T., Sep, S.J., van der Kallen, C.J., Dagnelie, P.C., Koster, A., Schaper, N., Henry, R., Stehouwer, C.D., 2014. The maastricht study: An extensive phenotyping study on determinants of type 2 diabetes, its complications and its comorbidities. Eur. J. Epidemiol. 29 (6), 439–451.
Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D., 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 618–626.
Sixt, L., Granz, M., Landgraf, T., 2020. When explanations lie: Why many modified bp attributions fail. In: International Conference on Machine Learning. PMLR, pp. 9046–9057.
Smilkov, D., Thorat, N., Kim, B., Viégas, F., Wattenberg, M., 2017. Smoothgrad: Removing noise by adding noise. arXiv preprint arXiv:1706.03825.
Snook, L., Plewes, C., Beaulieu, C., 2007. Voxel based versus region of interest analysis in diffusion tensor imaging of neurodevelopment. Neuroimage 34 (1), 243–252.
Springenberg, J.T., Dosovitskiy, A., Brox, T., Riedmiller, M., 2014. Striving for simplicity: The all convolutional net. arXiv preprint arXiv:1412.6806.
Sundararajan, M., Taly, A., Yan, Q., 2017. Axiomatic attribution for deep networks. In: International Conference on Machine Learning. PMLR, pp. 3319–3328.
Varol, E., Sotiras, A., Zeng, K., Davatzikos, C., 2018. Generative discriminative models for multivariate inference and statistical mapping in medical imaging. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 540–548.
Wachinger, C., Rieckmann, A., Pölsterl, S., Alzheimer's Disease Neuroimaging Initiative, et al., 2021. Detect and correct bias in multi-site neuroimaging datasets. Med. Image Anal. 67, 101879.
Walhovd, K.B., Fjell, A.M., Reinvang, I., Lundervold, A., Dale, A.M., Eilertsen, D.E., Quinn, B.T., Salat, D., Makris, N., Fischl, B., 2005. Effects of age on volumes of cortex, white matter and subcortical structures. Neurobiol. Aging 26 (9), 1261–1270.
Weichwald, S., Meyer, T., Özdenizci, O., Schölkopf, B., Ball, T., Grosse-Wentrup, M., 2015. Causal interpretation rules for encoding and decoding models in neuroimaging. Neuroimage 110, 48–59.
Wilming, R., Budding, C., Müller, K.-R., Haufe, S., 2022. Scrutinizing XAI using linear ground-truth data with suppressor variables. Mach. Learn. 1–21.
Wilms, M., Bannister, J.J., Mouches, P., MacDonald, M.E., Rajashekar, D., Langner, S., Forkert, N.D., 2022. Invertible modeling of bidirectional relationships in neuroimaging with normalizing flows: application to brain aging. IEEE Trans. Med. Imaging 41 (9), 2331–2347.
Worsley, K.J., Evans, A.C., Marrett, S., Neelin, P., 1992. A three-dimensional statistical analysis for CBF activation studies in human brain. J. Cereb. Blood Flow & Metab. 12 (6), 900–918.
Worsley, K.J., Friston, K.J., 1995. Analysis of fMRI time-series revisited—again. Neuroimage 2 (3), 173–181.

Wright, I., McGuire, P., Poline, J.-B., Travere, J., Murray, R., Frith, C., Frackowiak, R., Friston, K., 1995. A voxel-based method for the statistical analysis of gray and white matter density applied to schizophrenia. Neuroimage 2 (4), 244–252.

Xia, T., Chartsias, A., Wang, C., Tsaftaris, S.A., Alzheimer's Disease Neuroimaging Initiative, 2021. Learning to synthesise the ageing brain without longitudinal data. Med. Image Anal. 73, 102169.

Zeiler, M.D., Fergus, R., 2014. Visualizing and understanding convolutional networks. In: European Conference on Computer Vision. Springer, pp. 818–833.

Zhao, Q., Adeli, E., Honnorat, N., Leng, T., Pohl, K.M., 2019. Variational autoencoder for regression: Application to brain aging analysis. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 823–831.