NeuroImage 268 (2023) 119892

Contents lists available at ScienceDirect

NeuroImage

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

MC-RVAE: Multi-channel recurrent variational autoencoder for multimodal Alzheimer's disease progression modelling

Gerard Martí-Juan[a,*], Marco Lorenzi[b], Gemma Piella[a], for the Alzheimer's Disease Neuroimaging Initiative[1]

[a] BCN MedTech, Departament de Tecnologies de la Informació i les Comunicacions, Universitat Pompeu Fabra, Barcelona, Spain
[b] Université Côte d'Azur, Inria Sophia Antipolis, Epione Research Project, France

ARTICLE INFO

Keywords:
Alzheimer's disease
Longitudinal
Multimodal
Variational autoencoder
Recurrent neural network
Disease progression modelling

ABSTRACT

The progression of neurodegenerative diseases, such as Alzheimer's Disease, is the result of complex mechanisms interacting across multiple spatial and temporal scales. Understanding and predicting the longitudinal course of the disease requires harnessing the variability across different data modalities and time, which is extremely challenging. In this paper, we propose a model based on recurrent variational autoencoders that is able to capture cross-channel interactions between different modalities and model temporal information. These are achieved thanks to its multi-channel architecture and its shared latent variational space, parametrized with a recurrent neural network. We evaluate our model on both synthetic and real longitudinal datasets, the latter including imaging and non-imaging data, with $N = 897$ subjects. Results show that our multi-channel recurrent variational autoencoder outperforms a set of baselines (KNN, random forest, and group factor analysis) for the task of reconstructing missing modalities, reducing the mean absolute error by 5% (w.r.t. the best baseline) for both subcortical volumes and cortical thickness. Our model is robust to missing features within each modality and is able to generate realistic synthetic imaging biomarkers trajectories from cognitive scores.

1. Introduction

Alzheimer's Disease (AD) is an irreversible neurodegenerative disease that causes progressive cognitive impairment, leading to death (Lane et al., 2018). AD currently has no cure and causes an enormous strain on healthcare systems, due to the care needed by patients (Alzheimer's Association., 2018). AD is a multifactorial disease that affects different systems and processes of the brain (Jack et al., 2010). Those processes are captured using different biomarkers coming from various imaging modalities, such as magnetic resonance imaging (MRI) to capture atrophy, or positron emission tomography (PET) for glucose uptake or blood flow. Moreover, for a disease as complex as AD, it is necessary to study not only how those different biomarkers interact with each other, but also how they progress over time. For this reason, the use of multimodal longitudinal data, captured over several visits or time points, is extremely valuable to researchers. However, it is often difficult to use this kind of medical data, as it can present missing observations across modalities (e.g. impossibility of patients to receive invasive testing) and over time (e.g. due to medical follow-ups that were missed).

Combining information from imaging and non-imaging (e.g. genetics, clinical data) data and understanding how they relate to each other is crucial to enable an integrative analysis of the disease. For this task, commonly used methods are canonical correlation analysis (CCA) (Hardoon et al., 2004), which learns a shared representation between two modalities of data, partial least squares (PLS) (Édith Le Floch et al., 2012), which maximizes the covariance between two sets of latent variables, and reduced rank regression (RRR) (Vounou et al., 2010), among others. For more than two modalities, Group Factor Analysis (GFA) (Klami et al., 2015) presents a similar formulation introducing linear factors that describe the relationships between the modalities and the shared latent variables. More complex deep learning based methods have also been used to find a common representation for specific tasks, and to create models that are robust to missing

* Corresponding author.
E-mail address: gerard.marti@upf.edu (G. Martí-Juan).

[1] Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://www.adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf.

https://doi.org/10.1016/j.neuroimage.2023.119892.
Received 28 September 2021; Received in revised form 15 December 2022; Accepted 18 January 2023
Available online 20 January 2023.

information (Ngiam et al., 2011; Tsai et al., 2019). Recently, methods based on generative variational autoencoders (VAE) have been proposed to jointly learn a probabilistic latent space that encompasses the relation between different modalities (Shi et al., 2019; Wu and Goodman, 2018). Antelmi et al. (2019) proposed a multi-channel VAE that, using multiple modalities of cross-sectional data, is able to separate between diagnosis and show relationships between modalities. Their method infers a lower dimensional latent space that accounts for dependencies across the channels, being able to generate those that were missing in specific patients. These methods, however, have only been applied on datasets with no temporal information.

The problem of analysing longitudinal data can be tackled as a sequence learning problem. Cao et al. (2019) proposed a method using CCA to compare two different time series, and applied it to functional MRI analysis. Recurrent neural networks (RNN) (Hochreiter and Schmidhuber, 1997) have also become a powerful tool for sequence learning in the past few years. Although they have been mainly used for natural language processing (Goldberg, 2017), they have also been applied in other fields, such as image generation (Gregor et al., 2015) and AD progression and prediction (Lee et al., 2019; Mehdipour Ghazi et al., 2019; Nguyen et al., 2018; Wang et al., 2018). Variational RNN models (Chung et al., 2015; Fabius and van Amersfoort, 2015) add a variational approximation of the latent space to the RNN architecture, allowing it to capture more complex temporal variations by setting dependencies across time between the random latent variables.

Modelling data that are both longitudinal and multimodal is a difficult task because it requires modelling both temporal and cross-modality variability, as well as inter- and intra- subject variability (Verbeke et al., 2014). For neurodegenerative diseases, progression modelling methods can be used to leverage medical data to identify and predict the trajectories of biomarkers. This can lead to better characterization and understanding of the evolution of the disease (Martí-Juan et al., 2020; Oxtoby and Alexander, 2017).

There has been various methods presented in the literature for neurodegenerative disease progression modelling. Some approaches infer clinical data trajectories from brain imaging data (Lei et al., 2020), directly model the trajectories (Fisher et al., 2019), or account for irregularly sampled data points (Moore et al., 2019). Marinescu et al. (2019) proposed a spatiotemporal progression model on the cortex surface, finding finer progression details of atrophy progression. El-Sappagh et al. (2020, 2021) proposed an explainable multilayer method with time-series neuroimage data for disease classification. Other methods focus on grouping different imaging or non-imaging modalities. Event based models (Fonteijn et al., 2012; Young et al., 2014; 2015) aim to discover, for a set of multimodal biomarkers, the order in which they will degenerate. These methods were later extended in Young et al. (2018, 2021) for unsupervised subtyping and determining progression stage. Other remarkable approaches to the problem are multi-task learning models, which divide the problem into several tasks (per modality, or per subject) and aim to solve them together (Aksman et al., 2019; Nie et al., 2017). Finally, Gaussian processes (Hyun et al., 2016; Lorenzi et al., 2019; 2015) allow characterizing the uncertainty of the progression.

Most of the methods described in this section are limited in the number of modalities they can combine or in the amount of longitudinal data they use (cross-sectional data, or only short-term data). Moreover, it is not clear how to combine longitudinal modalities that have different amount of time points and/or capture patients at different stages of the disease. This is still an open problem in longitudinal data analysis for neurodegenerative diseases.

In this paper, we propose a multi-channel recurrent variational autoencoder (MC-RVAE) for multimodal longitudinal disease progression modelling. Similarly to Antelmi et al. (2019), we jointly model the dependencies across the channels of the network, but extend this rationale by adding a variational recurrent block to capture the temporal dependencies alongside every channel. This kind of formulation can account for longitudinal data with a variable number of time points and missing modalities. We show that the model captures the progression of the disease and the relationships between the modalities and is able to reconstruct missing modalities over time, even in the presence of missing data within modalities. We provide an interpretable representation of the evolution and show that the parameters learned by the model can be used to interpret the learned relationships between modalities. In addition, we conduct a sensitivity analysis to evaluate the importance of each individual feature in each modality for the reconstruction. We also present qualitative results of cross-modality reconstructions, showing synthetic brain data reconstructed from non-imaging data that is consistent with the progression of the disease.

2. Materials and methods

All code used to produce the pipeline and experiments described in this paper can be found in the repository of the project.[2]

2.1. MC-RVAE: Multi-channel recurrent variational autoencoder

Let $x = (x_1, \dots, x_T)$ be a sequence of multivariate samples defined over T time points, where each $x_t = (x_t^1, \dots, x_t^C)$ is a set of observations in C channels. Thus, x_t^c denotes the input at channel c and time t. To simplify the notation, we assume that all channels have the same number of time points, although this can be extended to uneven sampling of measurements in time. We base our model on the assumption that there are dependencies across time steps and across channels, given the multifactorial and progressive nature of AD. To capture the temporal dependencies, we use a variational RNN, similar to the one described in Chung et al. (2015), where we generate latent variables z_t at each t, which are conditioned on the RNN hidden state h_{t-1} of the RNN at the previous temporal step. For the inter-channel dependencies we assume, similarly to Antelmi et al. (2019), that all channels share the same latent space z_t, and that each channel is conditionally independent from the others given z_t. Conditional independence implies that all the modalities are completely defined by the common latent factor. This latent variable is common to all the modalities and thus represents their common variability, i.e. their interdependence. Noteworthy, the assumption of conditional independence also brings some useful statistical properties, since the derivation of the lower bound is more tractable, and allows for the reconstruction across modalities. Figure 1 summarizes a summary diagram of the model and its main characteristics.

We first describe the encoder and the decoder of our model for a specific channel c and time point t:

$$\begin{aligned} q(z_t|x^c_{\leq t}, z_{<t}) &= \mathcal{N}(\mu_{z_t}, \Sigma_{z_t}), \text{ where } [\mu_{z_t}, \Sigma_{z_t}] = \varphi^c_{enc}(x^c_t, h_{t-1}) \\ p(x^c_t|z_{\leq t}, x^c_{<t}) &= \mathcal{N}(\mu_{x^c_t}, \Sigma_{x^c_t}), \text{ where } [\mu_{x^c_t}, \Sigma_{x^c_t}] = \varphi^c_{dec}(z_t, h_{t-1}), \end{aligned} \tag{1}$$

where $x^c_{\leq t}$ represents the set of x^c at time points between 1 and t and likewise for $z_{<t}$ from 1 to $t-1$. The first line of Eq. (1) corresponds to the approximate posterior or encoder, defined by a normal distribution parametrized by φ^c_{enc} for each channel, which can be any differentiable function (for example, a neural network). The second line corresponds to the decoder, defined similarly to the encoder. They both depend on h_{t-1}. The hidden state is updated recurrently by:

$$h_t = f_\theta(z_t, h_{t-1}), \tag{2}$$

where f_θ can be parametrized by any recurrent architecture. We can then define the prior of z_t as:

$$z_t \sim \mathcal{N}(\mu_{0,t}, \Sigma_{0,t}), \text{ where } [\mu_{0,t}, \Sigma_{0,t}] = \varphi_{prior}(h_{t-1}). \tag{3}$$

Accounting for the dependencies described by the recurrent step in Eq. (2), for any time point $T > 1$ and channel c, the encoder and decoder

[2] https://www.github.com/GerardMJuan/RNN-VAE.

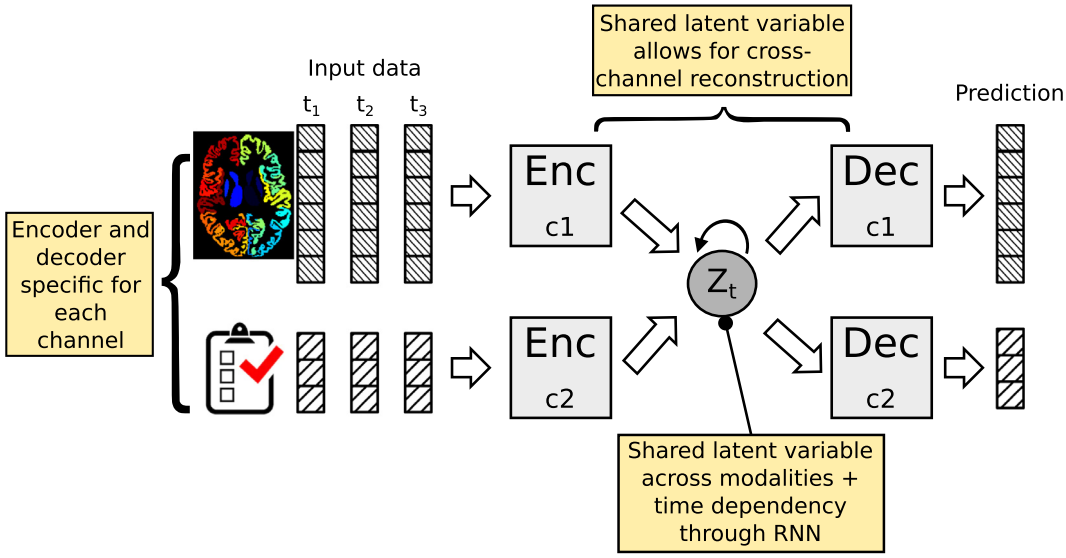

Fig. 1. Summary of the model. The input data is a sequence of modalities or channels (two in this figure), which can have different dimensionality and characteristics. Each channel has specific decoders and encoders, while the latent variable z, defined using a recurrent neural network, is shared across channels. This characteristic allows cross-channel reconstruction (for example, generating channel c1 from channel c2, or viceversa).

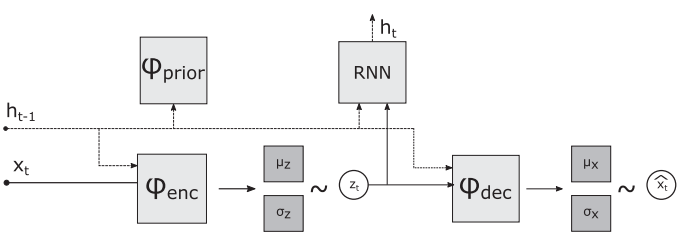

Fig. 2. Building block of the network, for a single channel, given an input x_t (channel index has been dropped for clarity). x_t is the input at time-step t, h_t is the hidden state of the RNN at time t, while h_{t-1} is the hidden state of the RNN at previous time-step $t-1$.

can be factorized as:

$$q(z_{\leq T}|x^c_{\leq T}) = \prod_{t=1}^{T} q(z_t|x^c_{\leq t}, z_{<t}) \tag{4}$$

$$p(x^c_{\leq T}, z_{\leq T}) = \prod_{t=1}^{T} p(x^c_t|z_{\leq t}, x^c_{<t})p(z_t|x_{<t}, z_{<t}). \tag{5}$$

These expressions will be used for the derivation of the lower bound of the model (Section 2.1.1). Figure 2 shows the main recurrent building block of the network, for a single channel. Figure 3 shows the recurrent and inference procedure in the network for multiple channels.

2.1.1. Lower bound

The optimization objective of the MC-RVAE is the evidence lower bound (ELBO) of the log-evidence of the data. This lower bound can be expressed as:

$$\mathcal{L} = \mathbb{E}_c\, \mathbb{E}_{q(z_{\leq T}|x^c_{\leq T})}\Big[\sum_{t=1}^{T}\Big(\sum_{i=1}^{C} \ln p(x^i_t|z_{\leq T}, x_{<t}) - \mathcal{D}_{KL}(q(z_t|x^c_{\leq t}, z_{<t})||p(z_t|x_{<t}, z_{<t}))\Big)\Big], \tag{6}$$

where $\mathcal{D}_{KL}$ denotes the Kullback-Leibler (KL) divergence. The first term of Eq. (6) forces the joint decoding of the channels at each time point, allowing us to predict missing channels. The second term, which acts as a regularization, forces the encoders of each channel to be close to a common prior generated from the previous time point, enforcing a temporal regularity. By maximizing this lower bound for the model parameters, we are also maximizing the log-evidence of the data. Supplementary file S1 contains the full derivation of the lower bound.

2.1.2. Practical implementation of MC-RVAE with heterogeneous data modalities

To take into account the different channels, we scale the reconstruction loss of each channel by the associated channel dimensions, rather than summing the losses up, as it is usually done when computing multivariate losses. In this way, we give the same importance to all channels regardless of their dimensionality.

Using a variable number of time points per subject (due to missing data) could lead to a bias towards samples with more time points. For this reason, we scale the lower bound in Eq. (6) by the number of non-missing time points to ensure that every subject contributes equally to the cost function regardless of their number of available time points.

With the differences in dimensionality across channels, dimensions in the shared latent space will encode different aspects of the channels: certain dimensions are shared across all channels, while others are used to represent with larger capacity channels that have larger dimensionality. For channels with a dimensionality smaller than z_{dim}, we could have an overfitting issue. To solve this, we introduce a constraint in the latent space so that we force the reconstruction from those channels to only use a subset of the latent space dimensions.

2.1.3. Variational dropout

Variational dropout was proposed by Kingma and Ba (2015) to regularize variational autoencoders and sparsify their latent space. This approach is based on defining a sparsity inducing prior on the latent weights, and on the parameterization of the variational posterior such that the estimated variance is associated to a "dropout" rate, which specifies the degree of sparsity associated with each latent dimension. This reparametrization is defined as:

$$q(z_t|x^c_{\leq t}, z_{<t}) = \mathcal{N}(\mu_{z_t}, \alpha\mu^2_{z_t}),$$
$$p(\mu_{z_t}) \propto \frac{1}{|\mu_{z_t}|}, \tag{7}$$

with α being the learned parameter shared across channels and time points, which quantifies the dropout rate for the given latent dimension. With this reparametrization, we can use the approximation of $\mathcal{D}_{KL}$ defined by Molchanov et al. (2017) (Eq. (8)) that only depends

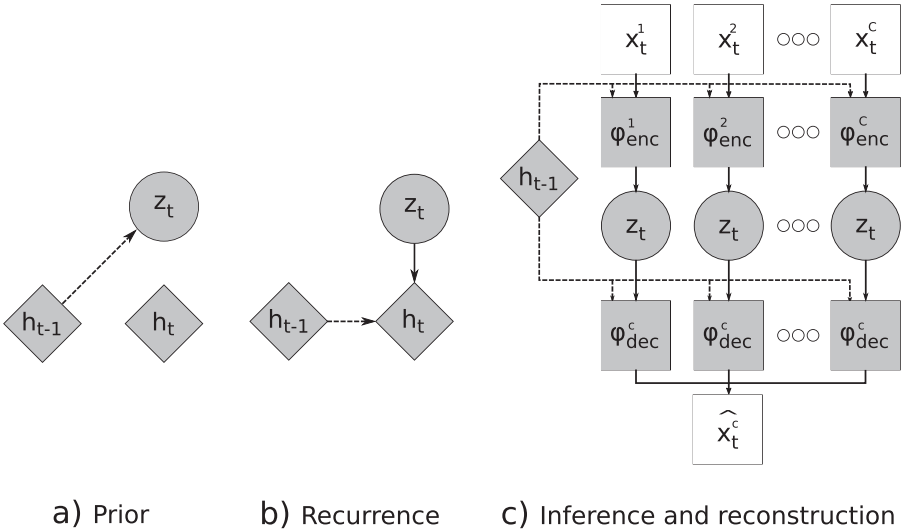

Fig. 3. Main processes of the model for inference and reconstruction. a) Prior of z_t, which depends on h_{t-1}. b) Recurrence step of the network, depending on h_{t-1} and z_t, and shared across channels. c) Inference process for a single channel c and a specific time point t: the decoder corresponding to channel c reconstructs the channel from the different computed latent projections encoded from each channel and the temporal information encoded in h_{t-1}. z_t represents the view of the latent space z at time t, encoded by each channel. $\hat{x}_t^c$ represents the predicted value.

on α (derivation of the approximation can be found in Molchanov et al., 2017).

$$\mathcal{D}_{KL}(q(z_t|x^c_{\leq t}, z_{<t})||p(\mu_{z_t})) \approx$$
$$-k_1\sigma(k_2 + k_3 \ln\alpha) + 0.5 \ln 1 + \alpha^{-1} + k_1, \text{with}$$
$$k_1 = 0.63576 \quad k_2 = 1.87320 \quad k_3 = 1.48695. \tag{8}$$

As in our model we have a learned prior for $t > 1$ (Eq. (3)), the sparsity constraint is applied uniquely at time $t = 1$ (the initial time point).

2.1.4. Reconstruction and inference

Thanks to the first term of the lower bound in Eq. (6) that forces the joint decoding of each channel, we can reconstruct the data for a missing channel using existing ones. For a missing channel c_m, a set of available channels c_a and for each time point t, we can decode the shared latent space z_t and reconstruct the missing channel using the following expression:

$$x_t^{c_m} = \mathbb{E}_{c_a}\mathbb{E}_{q(z_t|x_t^{c_m})}[\sum p(x_t^{c_a}|z_t)]. \tag{9}$$

2.1.5. Sensitivity analysis

To further interpret how the features in each modality or channel interact with each other, we performed a sensitivity analysis of the model. This was done by computing the Jacobian of the the trained model with respect to each input, defined as:

$$\mathbf{J}_{i,j} = \frac{\partial g_i(\mathbf{x})}{\partial x_j}, \tag{10}$$

where $J_{i,j}$ represents the partial derivative of the output $g_i(\mathbf{x})$ w.r.t. input x_j, with i and j being specific channels. Sensitivity analysis using the Jacobian has been shown to be comparable with sampling-based techniques in related works (Molamohammadi et al., 2020; Sobol' and Kucherenko, 2009). Sensitivity analysis for each modality and feature was done by measuring the mean of the square Jacobian matrix over all the subjects. For each output channel, we compute the Jacobian w.r.t each of the other input channels. This allows us to find the most important features in a specific channel for reconstruction of a different one and it gives us a clearer picture of the learned cross-channel relationships. We also averaged separately the Jacobians that correspond to Cognitively Normal (CN), Alzheimer's Disease subjects (AD) or patients with Mild Cognitive Impairment (MCI), to assess any differences across groups.

2.2. Data

In this section we present the two different datasets we use in our experiments: a generated synthetic dataset and a real medical dataset.

2.2.1. Synthetic data

We evaluated our method on a longitudinal synthetic dataset, which was generated from a set of latent variables of fixed dimensionality, as described in Eq. (11):

$$\begin{aligned} &z, z_v \sim \mathcal{N}(0, I_{z_{dim}}) \\ &\varepsilon \sim \mathcal{N}(0, I_g) \\ &x_t^c = W^c(z + z_v \times \lambda \times t) + \varepsilon \end{aligned} \tag{11}$$

$I_{z_{dim}}$ is the identity matrix of size $z_{dim} \times z_{dim}$, where z_{dim} is the dimensionality of the latent space. $W^c \in \mathbb{R}^{g \times z_{dim}}$, with g being the dimensionality of the output channel c, is a random orthonormal matrix that linearly relates each channel c to the shared latent space, and ε is Gaussian noise. Longitudinal samples are generated by a random linear transformation through the latent space via a variable z_v for t time points, with this translation modulated by a scalar λ. The number of time points is variable for each generated sample, with the number of time points ranging from 5 to 11. Supplementary Figure S2 shows a projection of the latent data and examples of trajectories over time.

2.2.2. ADNI data

Data used in this paper were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (Mueller et al., 2005). The primary goal of ADNI has been to study whether serial imaging and biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of MCI and early AD. Informed consent was obtained from all individual participants in the study by ADNI. Restrictions apply to the availability of these data. Data are available at https://adni.loni.usc.edu/.

We focused on three different data modalities (i.e. channels). Full information on the features used can be found in supplementary file S3:

- Brain subcortical volumes: We selected MRI T1 weighted images, preprocessed using gradient warping, scaling, B1 correction, and N3 inhomogeneity correction. Images were then automatically processed with the longitudinal pipeline (Reuter et al., 2012) in FreeSurfer. Specifically, an unbiased within-subject template space and image was created using robust, inverse consistent registration (Reuter et al., 2010). Several processing steps, such as skull stripping, Talairach transform, atlas registration as well as spherical surface maps and parcellations were then initialized with common information from the within-subject template, significantly increasing

Table 1
Demographic characteristics of the cohort at baseline. Age and education presented as average and standard deviation, in years. N: Number of subjects. APOE $\varepsilon 4$: Apolipoprotein $\varepsilon 4$, percentage with 1 or 2 alleles. CN: Cognitively normal. MCI: Mild cognitive impairment. AD: Alzheimer's disease. MMSE: Mini-mental state examination.

	CN	MCI	AD	Total
N	229	324	180	897
Age	74.6±5.3	73.7±7.3	73.7±7.8	73.6±7.0
Education	16.2±2.7	15.7±3.0	15.0±2.9	15.8±2.9
MMSE	29.06±1.09	27.13±1.80	23.19±1.97	27.09±2.65
Female	51.53%	40.12%	50.55%	46.70%
APOE $\varepsilon 4$	27.51%	56.17%	51.10%	47.80%

Table 2
Hyperparameters used in the model trained using synthetic and ADNI data. Left column: parameters used with the model trained on synthetic data. Right column: parameters used for the model trained on ADNI data. Layer size: size of all layers in the network (encoder, decoder and φ). RNN hidden: size of the hidden layer in the RNN. Parameters selected for the final model are highlighted.

Hyperparameters	Synthetic	ADNI
No. layers	**0**, 1	**0**, 1, 2, 3
Layer size	10, 20	20, 50, 80
No. latent dim.	**20**, 50, 100	15, 20, **30**
RNN hidden size	**5**, 10, 30	30, **50**, 70

reliability and statistical power. We used the parcelled subcortical volumes, obtaining 40 features.

- Cortical thickness: We used the same imaging pipeline described above and obtained the parcelled cortical thickness, obtaining 68 features.
- Cognitive assessment scores: A set of six different neuropsychological assessment tests that capture the level of cognitive decline of patients in specific tasks and domains.

We selected subjects that had no missing data at their baseline acquisition for any of the channels, and all subsequent follow-ups that had at least one of the existing longitudinal modalities and had no missing features for this modality. With these criteria, we selected 897 subjects, with a total of 7224 multimodal acquisitions. Among those acquisitions, we extracted 3799 subcortical volumes, 3813 cortical, and 4847 cognitive score measures. This allows us to train and test our model with missing modalities and variable number of follow-ups.

Table 1 shows the demographic information of the selected cohort. Regarding the distribution of the data across modalities, supplementary Figure S4 shows the information about missing, mean and maximum number of acquisitions across modalities. Median time interval between acquisitions is 6.±1 months (Median ± Interquartile range). The number of acquisitions per diagnosis is 5 ± 1 for AD, 7 ± 10 for CN and 8 ± 4 for MCI.

Other relevant modalities were considered, such as PET derived biomarkers or cerebrospinal fluid biomarkers, but were ultimately not included because of their low amount of subjects with acquisitions available in the cohort and the low number of follow-ups.

2.3. Experiments

For our synthetic experiments, we generated 1000 samples, using 3 channels and 15 features per channel, with time points for each sample ranging from 5 to 11. The parameters used to train the model for these data are described in Table 2 (first column). For the ADNI cohort, we first separated our data between training (80%) and test (20%) set, with 717 and 180 subjects respectively. We stratified data by diagnosis across sets. The test set was then held out, and we performed a grid search on the training set over a set of hyperparameters, shown in Table 2 (second column). Parameters selected for the final model are highlighted. With no extra layers, φ_{enc} and φ_{dec} (Fig. 2) are parametrized as linear transformations, so the only non-linearity of the network is the recurrent step. We also applied the constraint described in Section 2.1.2 on the number of dimensions used in the latent space for the cognitive channel, limiting its dimensionality to 5, so that the latent dimension is lower than the actual dimension of the channel.

The model was trained using gradient descent with Adam optimization(Kingma and Ba, 2015), at a learning rate of 2e-3 until convergence (roughly 3000 epochs). Code and training settings of the model can be found in the repository of the paper.

We evaluated the results on the held-out test set on two reconstruction tasks using the joint decoder approach described in Section 2.1.4:

- **Cross-channel reconstruction:**, where each different channel was reconstructed from each of the other channels.
- **Full channel reconstruction:**, where each channel was reconstructed from the rest of the channels.

We evaluated both tasks using mean absolute error (MAE).

To test the robustness of MC-RVAE to missing features within modalities, the previous tasks were also performed while adding increasing amounts of random missing features for each channel and subject in the test set.

To assess the temporal prediction of our model, we removed the last time points from each subject in the test set, and we predicted them using the rest of the data. We tested the prediction capabilities of our model using one, two and three time points. We also did an additional experiment including two non-longitudinal modalities, with only information on the first acquisition: demographic information and APOE. For those channels we did not perform the recurrent step, and the information of the channel present at $t = 1$ is reproduced at all time points for calculating the joint reconstruction across channels.

2.3.1. Baseline methods

We compared our models to three different baseline methods:

- K nearest neighbors (KNN) imputation: for each sample (subject data at a given time point) in the test set, we performed a KNN search for each channel c to find the k most similar samples in our training set (in all our experiments $k = 5$). We use the average (along c) of these similar samples as an estimation of channel c in the test sample. When having missing features within a modality, we used an additional KNN to impute those values, per channel.
- Random forest prediction: we computed a linear model describing the temporal progression for each sample and channel in the training set. Then, we trained a random forest regressor model to predict the estimated linear progression parameters (intercept and slope) channel-wise. We used this model to predict the parameters of a linear progression for each test sample, and used this linear model to make the prediction. This baseline only allows us to perform reconstruction from one channel to another.
- Group factor analysis (GFA): we used GFA (Klami et al., 2015) for missing channel reconstruction. GFA generates linear factors describing the relationships across channels, and is able to model two or more channels. Implementation was done with R, using the default parameters.

Results were compared using an ANOVA model for each task. Differences across methods were evaluated with post hoc pairwise Tukey HSD tests.

2.3.2. Qualitative experiments

To illustrate the generative capabilities of our approach, we reconstructed cortical and subcortical data of two synthetic patients: one with

a stable, healthy cognitive trajectory, and another one that is rapidly declining. We generated four time points for each of the subjects and visualized the longitudinal neurodegeneration trajectories and their plausibility. The longitudinal synthetic data generated can be found in Supplementary Data S5. We compared these trajectories to cortical and subcortical trajectories obtained by averaging the first four time points from CN and AD subjects from the training set.

3. Results

3.1. Synthetic results

Table 3 shows the quantitative reconstruction results from a separate test set, compared to the baseline methods.

3.2. ADNI results

Table 4 shows the MAE of the cross-channel (top) and full channel (bottom) reconstruction, on the hold out ADNI test set, for our model and the baseline models for comparison. Figure 4 shows the performance of the model and KNN for increasing amounts of missing data within each modality. The MAE obtained by KNN imputation is also shown for comparison. Results on the prediction task are found on Supplementary file S6, and the results with the two additional non-longitudinal channels (demographics and APOE) can be found in Supplementary file S7.

Figure 5 shows the latent space obtained by MC-RVAE on the test set coloured by diagnosis, age, and time point, for the two latent dimensions with lowest dropout (dropout ratio of each latent dimension can be found in Supplementary Figure S8).

Figure 6 shows the weights of φ_{enc} for the three different channels (cognitive scores, cortical thickness and subcortical volumes) for the two latent space dimensions represented in Fig. 5, plus another latent space dimension that was selected by variational dropout, but did not have weights associated to cognitive scores (see Section 2.1.2). Regions and features for each modality (columns) are colored blue if they are directly correlated to the corresponding dimension (row), or red if inversely correlated. Fig. 7 shows the cross-channel sensitivity analysis results on the model. Sensitivity analysis by disease stage are shown in Supplementary Figure S9.

Figure 8 shows four synthetic time points of cortical and subcortical data. We sampled them using two sequences of 4 time points of cognitive data: a typical sequence of healthy control and a sequence of a rapidly declining patient, with similar demographic information. The colour scale displayed on the figure was obtained by computing the negative z-score of the generated data with respect to a control group of cognitively healthy, amyloid-negative patients (Shaw et al., 2009).

4. Discussion

We propose a generative model based on recurrent variational autoencoders that is able to jointly model a latent trajectory from multimodal, longitudinal data. We introduce the main concepts and assumptions behind the model, derive its lower bound, and test it both on synthetic and real data from a cohort of patients afflicted by AD to reconstruct missing modalities by leveraging the cross-channel information learned by the model.

For the synthetic data, we show the strength of our model to reconstruct longitudinal signals, compared to the baseline methods (Table 3). Our model is able to recover the latent trajectory of the data to reconstruct missing channels, even when no information about the original channel is fed to the model. The baseline methods do not leverage the longitudinal data and treat each time point as a separate unit, whereas MC-RVAE is able to capture those trajectories across patients and uses this variability to improve the prediction across channels. However, this can come at a cost, as our within-channel reconstruction results are subpar for some cases (C2 reconstruction). The latent representation is op-

Table 3
Cross-channel reconstruction (top table) and full channel reconstruction (bottom table) results, for the synthetic data experiments. The different synthetic channels are named C1, C2 and C3 for distinction. Rows are the original channels, columns the target channel. Statistical significance is defined with a post hoc Tukey HSD test between the best result and the second best result after an ANOVA test across all methods. KNN: K nearest neighbors. RF: random forest. GFA: Group factor analysis. MC-RVAE: multi-channel recurrent variational autoencoder. All values are mean absolute error (MAE) and standard deviation of the error over each subject and time point. Best results are highlighted. $^{*}p < 0.05$. $^{**}p < 0.01$. $^{***}p < 0.001$.

	KNN			RF			GFA			MC-RVAE		
	C1	C2	C3	C1	C2	C3	C1	C2	C3	C1	C2	C3
C1	0.38 ± 0.13	0.44 ± 0.14	0.5 ± 0.14	0.73 ± 0.25	0.72 ± 0.25	0.78 ± 0.25	-	0.67 ± 0.2	0.69 ± 0.21	**0**.22 ± 0.11***	**0**.29 ± 0.15***	**0**.22 ± 0.1***
C2	0.44 ± 0.13	0.36 ± 0.11	0.44 ± 0.12	0.77 ± 0.28	0.7 ± 0.24	0.75 ± 0.26	0.69 ± 0.2	-	0.68 ± 0.2	**0**.36 ± 0.17***	**0**.33 ± 0.17***	**0**.32 ± 0.15***
C3	0.48 ± 0.13	0.41 ± 0.12	0.41 ± 0.12	0.8 ± 0.26	0.74 ± 0.26	0.77 ± 0.26	0.7 ± 0.21	0.67 ± 0.2	-	**0**.14 ± 0.08***	**0**.21 ± 0.14***	**0**.02 ± 0.01***

	KNN	GFA	MC-RVAE
C1	0.44 ± 0.13	0.49 ± 0.12	**0**.24 ± 0.11***
C2	0.39 ± 0.11	0.44 ± 0.11	**0**.33 ± 0.16***
C3	0.44 ± 0.12	0.48 ± 0.12	**0**.14 ± 0.08***

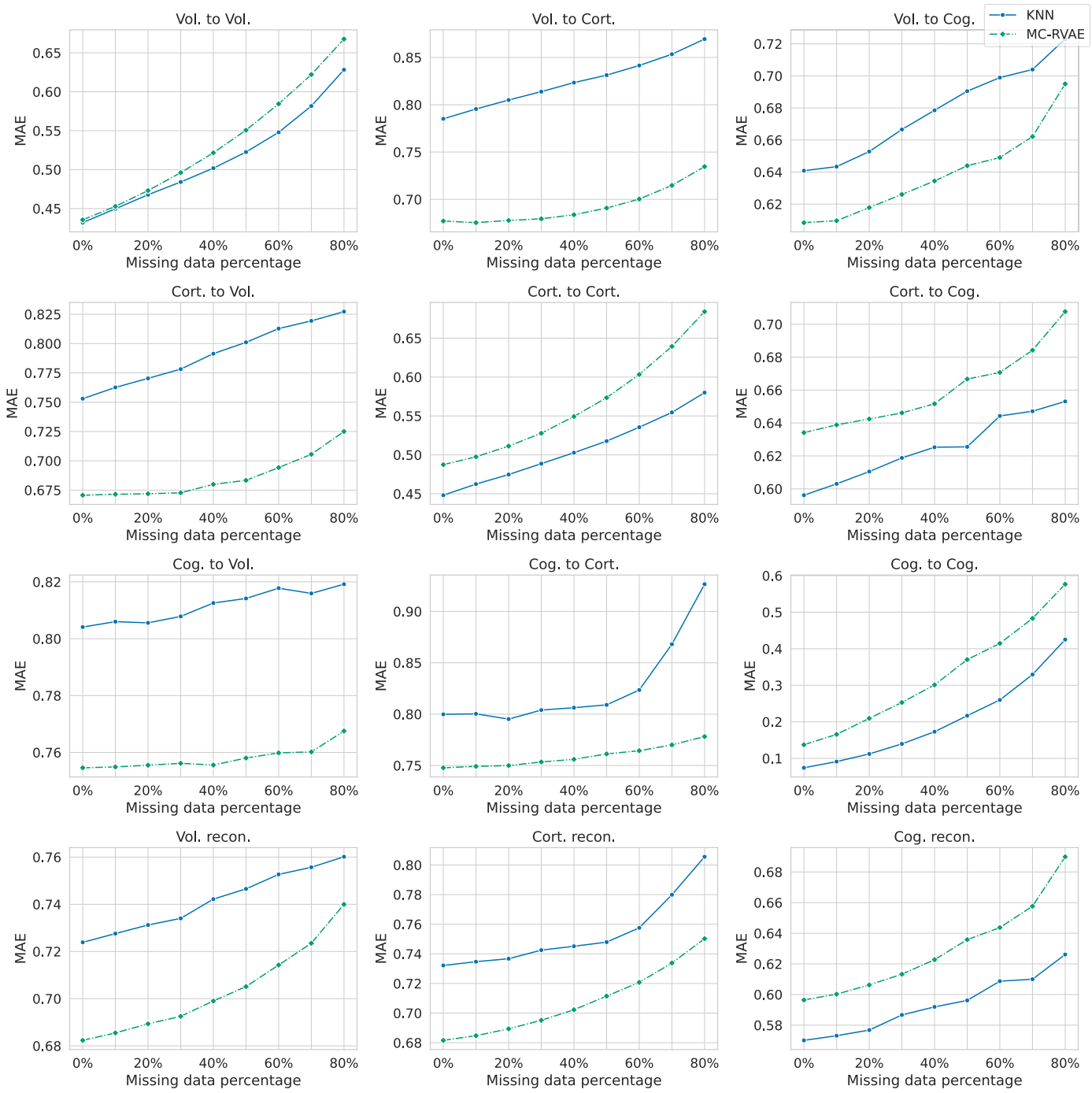

Fig. 4. Reconstruction performance with missing data. The first three rows correspond to the cross-reconstruction task across modalities, and the last row to the full reconstruction task.

timized to solve all the reconstruction problems jointly, and thus there is not a privileged modality to reconstruct in the cost function. This does not happen for a single-modality model, which lacks the cross-modality capability.

For the ADNI cohort, we compare the performance of our model on reconstructing each channel from the others, both individually and jointly. For the first task, the performance of our model is comparable to GFA and KNN, with the latter slightly outperforming our model in terms of MAE. However, our results, while having an overall slightly worse MAE, have a much lower standard deviation (Vol. to Cog. reconstruction KNN vs MC-RVAE, for example, Table 4), meaning that the predictions are more stable. Moreover, the better performance of KNN is mainly on the cognitive score prediction task, where, given the nature of the features (values situated on a narrow specific range), a nearest neighbors approach would naturally have lower errors, as the predictions would always be on a "valid" value. From the computational cost perspective, MC-RVAE is faster than the KNN baseline at inference time (0.1s compared to $\sim 1s$), although MC-RVAE uses GPU acceleration.

For full channel reconstruction, our model outperforms the baselines for volume and cortical thickness, although it has a similar performance for cognitive scores, likely due to the same reasons pointed before. MC-RVAE is robust to missing data within a channel, specially in the cross-channel reconstruction task (volume to cortical thickness, or cortical

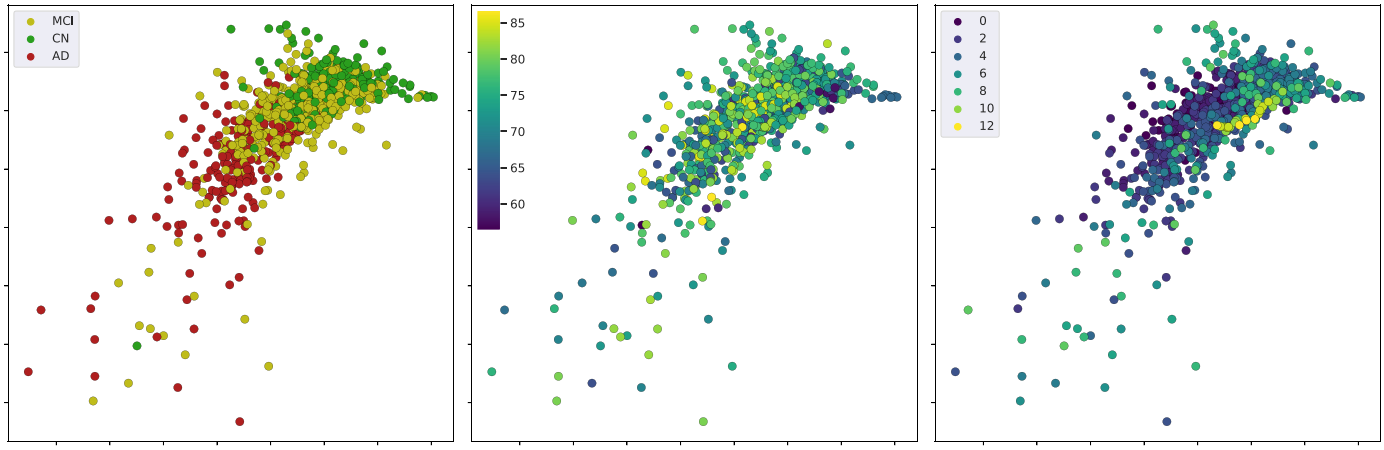

Fig. 5. Visualization of top two selected dimensions from latent space by variational dropout, for subjects on the hold-out test set. Each point represents a different time point of a subject. Left: colored by diagnosis. Middle: colored by age. Right: colored by time point.

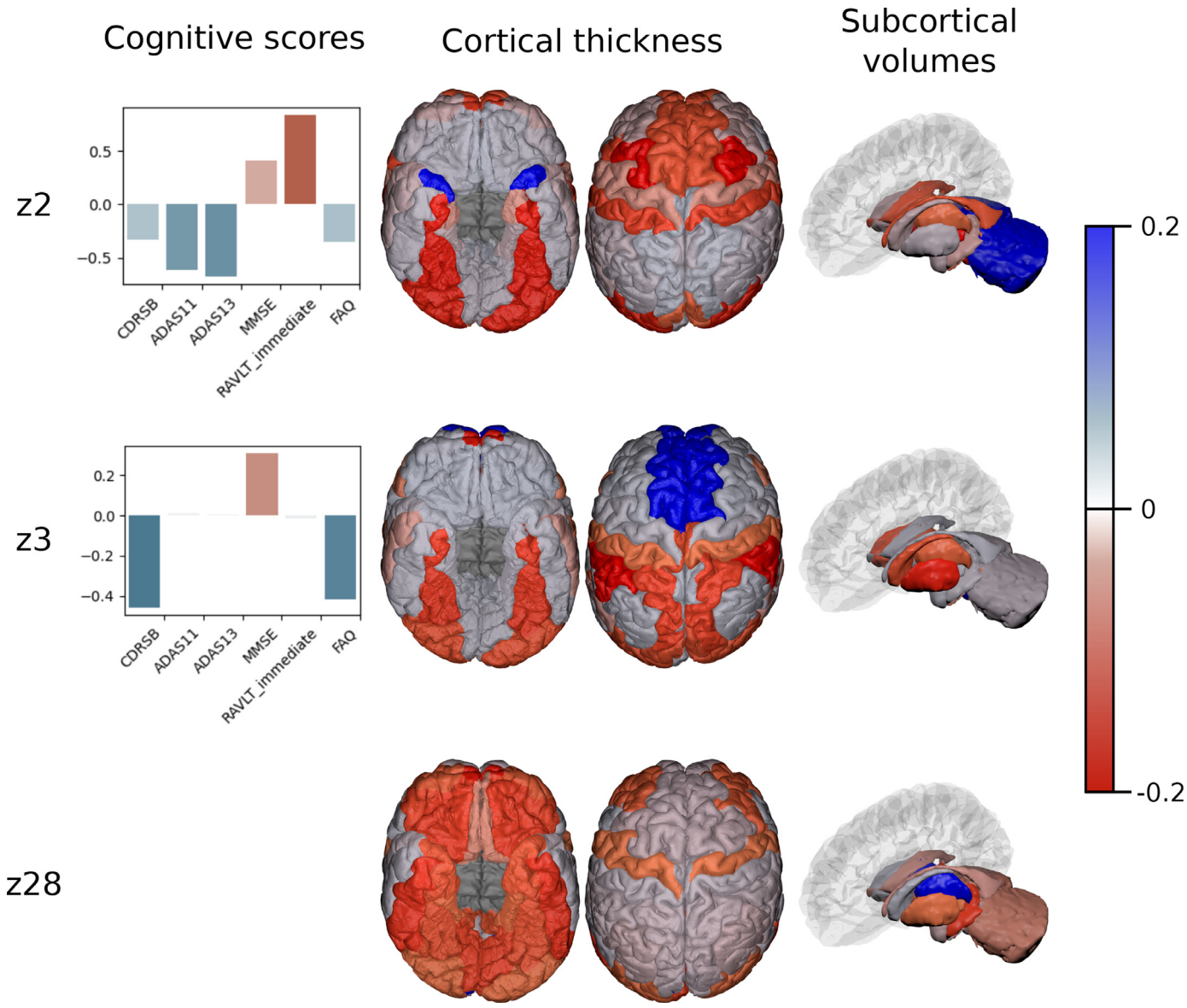

Fig. 6. Weights encoded by φ_{enc}, for the two latent dimensions represented in Fig. 5, plus the next dimension with lowest dropout. Note that z_{28} does not have weights associated with cognitive scores, hence the missing graphic. Brain figures generated using Brainpainter (Marinescu et al., 2019).

Table 4
Cross-channel (top) and full channel (bottom) reconstruction results, for the ADNI data. All results computed over test set. All values are mean absolute error (MAE) and standard deviation of the error over each subject and time point. Best results are highlighted. Rows are the original channels, columns the target channel. Statistical significance is defined with a post hoc Tukey HSD test between the best result and the second best result after an ANOVA test across all methods. Vol.: MRI subcortical volumes. Cort.: MRI cortical thickness values. Cog.: cognitive values. KNN: K nearest neighbors. RF: random forest. GFA: Group factor analysis. MC-RVAE: multi-channel recurrent variational autoencoder. $^{***}p < 0.001$.

	KNN			RF			GFA			MC-RVAE		
	Vol.	Cort.	Cog.	Vol.	Cort.	Cog.	Vol.	Cort.	Cog.	Vol.	Cort.	Cog.
Vol.	**0**.43±0.09	0.79±0.23	**0**.64±0.44***	0.74±0.18	0.96±0.34	0.95±0.82	-	0.78±0.28	0.71±0.4	0.44±0.09	**0**.67±0.16***	0.75±0.19
Cort.	0.75±0.19	**0**.45±0.08***	**0**.6±0.39***	0.99±0.26	0.81±0.22	0.93±0.84	0.74±0.2	-	0.7±0.39	**0**.68±0.19***	0.49±0.09	0.75±0.26
Cog	0.8±0.21	0.8±0.27	**0.07±0.06*****	1.06±0.37	1.11±0.43	0.44±0.37	0.76±0.2	0.79±0.29	-	**0**.61±0.51***	**0**.63±0.51***	0.14±0.08

	KNN	GFA	MC-RVAE
Vol.	0.72±0.18	0.74±0.20	**0**.68±0.16***
Cort.	0.73±0.22	0.77±0.28	**0**.68±0.21***
Cog.	**0**.57±0.38	0.68±0.38	0.59±0.48

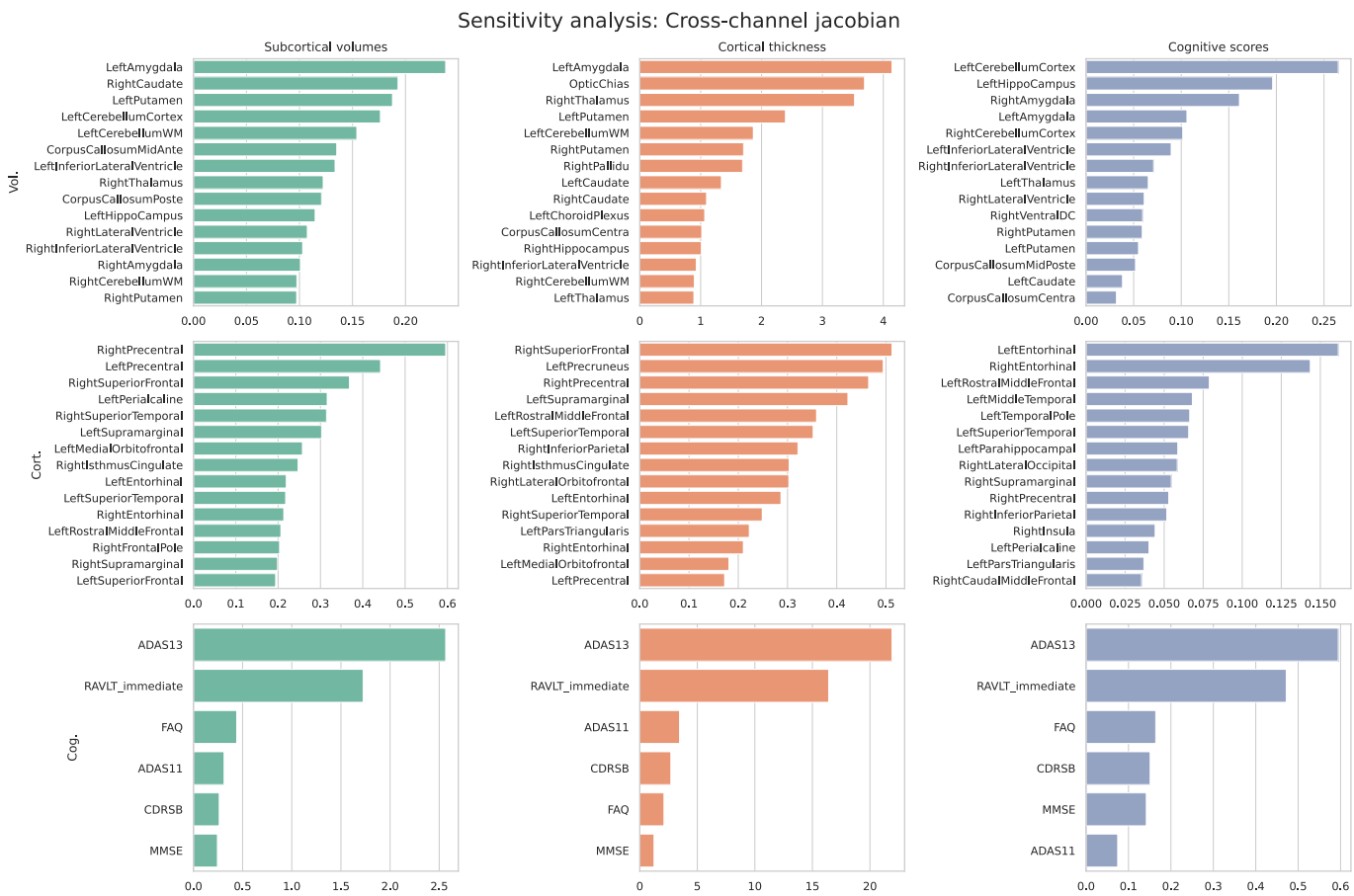

Fig. 7. Cross-channel sensitivity analysis. Each column corresponds to a different reconstruction target (also colored), with each bar representing the importance of that feature for reconstructing that modality.

thickness to volume), thereby demonstrating its ability for missing data inference.

Variational dropout selects 16 (out of 30) dimensions of z, including the subset of latent dimensions that were selected by constraining the latent space for the cognitive score channel due to its lower dimensionality (see Section 2.1.2). We observe a good separation between diagnosis and age, but we do not find a distinct temporal structure. Moreover, subjects with AD have less time points. This might bias our results, since temporal trajectories at later stages of the disease are not as well represented in our data as trajectories at earlier and middle stages.

As for the hyperparameters of our model, we tried using a deeper network, but no significant improvements were achieved (Supplementary Figure S10). This could be due to the noise present in this type of neuroimaging data over time; the overall signal is already captured by the simpler network, and adding more complexity to the model leads to overfitting. Another reason could be the sample size. More data and time points per subject could help unraveling the temporal and multimodal relations of the data, and in that case, a deeper model could boost the performance.

Since φ_{enc} and φ_{dec} (Eq. (1)) are linear functions, their parameters tell us how the network encodes the information to the shared latent

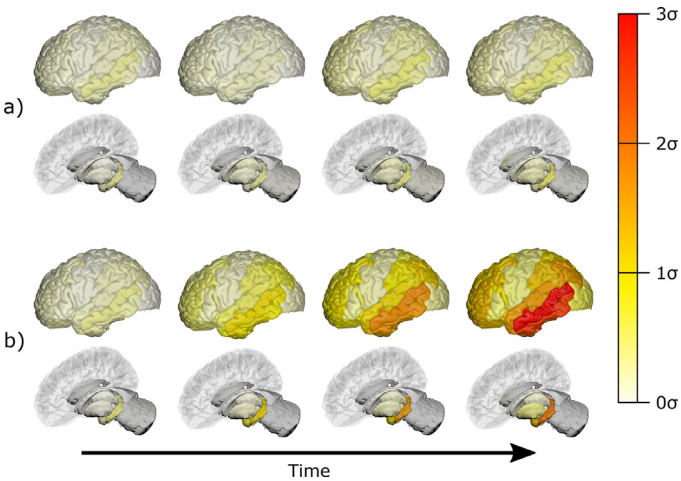

Fig. 8. Trajectories of synthetic cortical and subcortical generated by MC-RVAE. σ: standard deviation with respect to the control population. a) and b) respectively corresponds to simulated imaging progressions associated to stable and pathological cognitive progressions (Section 2.3.2). Figures were generated using Brainpainter (Marinescu et al., 2019).

space. For z_2, the dimension is strongly related to cognitive health, with MMSE and RAVLT being directly related to healthy cognitive ability (and thus have positive weights) and the other ones being inversely related (with negative weights). This relationship can also be seen in Fig. 5 (left). Regarding cortical areas, we observe large weights associated to middle and superior frontal areas, motor cortex, central sulcus and cuneus, whereas subcortical areas show importance in the caudate and amygdala. For z_3, we observe a higher importance of CDRSB and FAQ for the cognitive scores, and a large negative weight on the middel frontal lobe compared to z_2. On subcortical volumes, the network assigns a larger relevance to the cerebellum cortex and the amygdala, and none to the caudate. Temporal lobes and the hippocampi are not given importance for those two latent dimensions, which means that such variation is captured in other dimensions, and the network prioritizes cognitive changes to capture the disease (given the large weights associated to cognitive values in z_2 and the differentiation between diagnosis in Fig. 5, left). Looking at z_{28}, which is a dimension only characterized by cortical and subcortical volumes, strengths this hypothesis. We can observe large weights associated to the cortex, in a general, symmetric way, and with a large importance to subcortical areas such as the hippocampus.

Sensitivity analysis reveals the contribution of each feature for the cross-channel reconstruction task. For reconstructing subcortical volumes (green column), the most important cortical regions were the precentral (right and left), while ADAS13 and RAVLT were the cognitive tests with highest relevance, something also observed with the other channels. When trying to predict cognitive scores (blue column), cerebellum, hippocampus, amygdala, thalamus and ventricles are the most relevant subcortical regions. Atrophy on those regions is associated with an impact on cognitive functions (Jacobs et al., 2017; Yi et al., 2016). For the cortex, the most important regions to predict cognitive scores were the entorhinal, which has also been associated with cognitive impairment in AD (Du et al., 2001). This further reinforces the idea that MC-RVAE learns relevant associations across modalities to perform the cross-reconstruction task. No large differences were observed when separating the analysis by disease stage, only a higher influence of cognitive scores (specially for ADAS13 and RAVLT, but also for the rest) for AD and MCI subjects compared to CN, which might indicate that the model has learned to focus on cognition to predict the evolution of the imaging biomarkers when the patient is in a worse state.

Performance of the prediction tasks was not worse than the one obtained by our baseline methods (constant value and a linear model). The experiment adding two non-longitudinal channels showed the model improving in some areas while underperforming in others (same channel reconstruction, for example). Overall, results in those experiments are similar to those obtained by the baseline methods, motivating further extensions of the current approach to jointly account for longitudinal and non-longitudinal data.

To further visualize and interpret the reconstruction capability of the model and how the channels are related to each other, Fig. 8 shows two different sequences of cortical and subcortical MRI data generated by the network, for a cognitive healthy patient and a rapidly declining patient. The generated data present a plausible decline of cortical thickness and subcortical volume for both cases, remaining consistent alongside the individual trajectory of the subject. Subject a) shows stability across time for a stable cognitive trajectory, whereas subject b) shows a rapid decline on the hippocampus and across the cortex, especially in the temporal cortex, which are areas known to be directly affected by AD (Bakkour et al., 2013). Compared to the sequences obtained by averaging real CN and AD data (Supplementary Figure S11), the synthetic trajectories are close, with the main difference being that the model predicts a steeper atrophy at later time points compared to the average trajectories. Averaging data, however, has a smoothing effect that could explain this difference.

Compared to existing methods, our model principal contribution is its flexibility to scale to a larger number of channels and a variable number of time points. For example, the model proposed in Cao et al. (2019), based on canonical correlation analysis, can only use two different longitudinal channels and is not generative. Compared to the multi-channel model of Antelmi et al. (2019), from which our model is built upon, our implementation includes temporal structure and allows for temporal data to be included and generated.

We developed and tested a generative model on multimodal, longitudinal data based on recurrent variational autoencoders. The model can combine different modalities of variable dimension acquired across time, learn their cross channel dependencies and intra-channel temporal dependencies, and use this information to reconstruct and predict missing modalities. Results on synthetic and real medical data show strong performance on missing data reconstruction, suggesting its potential for future time point prediction. A hindrance of our model is the assumption of equal time spacing between acquisitions. In our (real) dataset, spacing between time points has a small variance, so the impact of this assumption should be minor. However, specifying temporal information could allow the model to sample synthetic data at specific future points, which could be of use for clinical trial design.

Moreover, we noticed that variational dropout tends to select unnecessary latent dimensions and does not fully sparsify the latent space, so a better approach to introduce this concept and its assumptions to the model would improve its performance and interpretability. Finally, we should take advantage of the uncertainty learnt by the model to understand for which tasks and situations the model has less confidence, and improve on it. The reconstruction capabilities of the proposed model and its ability to integrate longitudinal information and relate imaging and non-imaging data can be useful to further understand the temporal evolution of the disease and, in a more practical way, to generate patient coherent missing scans.

Data and Code Availability Statement

All code used to produce the pipeline and experiments described in this paper can be found in the repository of the project https://www.github.com/GerardMJuan/RNN-VAE. Data used in this paper were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database [Mueller et al, 2005]. The primary goal of ADNI has been to study whether serial imaging and biological markers, and clinical and neuropsychological assessment can be combined to

measure the progression of MCI and early AD. Informed consent was obtained from all individual participants in the study by ADNI. Restrictions apply to the availability of these data. Data are available at http://www.adni.loni.usc.edu-upon-application.

Declaration of Competing Interest

None.

Credit authorship contribution statement

Gerard Martí-Juan: Methodology, Software, Validation, Formal analysis, Investigation, Writing – original draft. **Marco Lorenzi:** Conceptualization, Formal analysis, Resources, Supervision, Writing – review & editing. **Gemma Piella:** Supervision, Resources, Project administration, Funding acquisition, Writing – original draft.

Data availability

Data will be made available on request.

Acknowledgments

This work is supported by the European Union's Horizon 2020 research and innovation programme (grant n° 848158). M. Lorenzi is supported by the French government, through the 3IA Côte d'Azur Investments in the Future project managed by the National Research Agency (ANR) (ANR-19-P3IA-0002). G. Piella is supported by ICREA under the ICREA Academia programme. This publication is part of the project PCI2021-122044-2A, funded by the project ERA-NET NEURON Cofund2, by MCIN/AEI/10.13039/501100011033/ and by the European Union "NextGenerationEU"/PRTR. Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (http://www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

Supplementary material

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.neuroimage.2023.119892.

References

Aksman, L.M., Scelsi, M.A., Marquand, A.F., Alexander, D.C., Ourselin, S., Altmann, A., 2019. Modeling longitudinal imaging biomarkers with parametric Bayesian multi-task learning. Hum. Brain Mapp. 40 (13), 3982–4000. doi:10.1002/hbm.24682.

Alzheimer's Association., 2018. Alzheimer's Dement.: Global Resources. https://www.alz.org/global/.

Antelmi, L., Ayache, N., Robert, P., Lorenzi, M., 2019. Sparse multi-channel variational autoencoder for the joint analysis of heterogeneous data. In: 36th Int. Conf. Mach. Learn., ICML 2019, pp. 453–464.

Bakkour, A., Morris, J.C., Wolk, D.A., Dickerson, B.C., 2013. The effects of aging and Alzheimer's disease on cerebral cortical anatomy: specificity and differential relationships with cognition. Neuroimage 76, 332–344. doi:10.1016/j.neuroimage.2013.02.059.

Cao, X., Ke, J., Sandstede, B., Luo, X., 2019. Time-dependent canonical correlation analysis for multilevel time series. bioRxiv doi:10.1101/650101.

Chung, J., Kastner, K., Dinh, L., Goel, K., Courville, A., Bengio, Y., 2015. A recurrent latent variable model for sequential data. Adv. Neural Inf. Process. Syst. 2980–2988.

Du, A.T., Schuff, N., Amend, D., Laakso, M.P., Hsu, Y.Y., Jagust, W.J., Yaffe, K., Kramer, J.H., Reed, B., Norman, D., Chui, H.C., Weiner, M.W., 2001. Magnetic resonance imaging of the entorhinal cortex and hippocampus in mild cognitive impairment and alzheimer's disease. J. Neurol. Neurosurg. Psychiatry 71 (4), 441–447. doi:10.1136/jnnp.71.4.441.

El-Sappagh, S., Abuhmed, T., Riazul Islam, S.M., Kwak, K.S., 2020. Multimodal multitask deep learning model for Alzheimer's disease progression detection based on time series data. Neurocomputing 412, 197–215. doi:10.1016/j.neucom.2020.05.087.

El-Sappagh, S., Alonso, J.M., Islam, S.M., Sultan, A.M., Kwak, K.S., 2021. A multilayer multimodal detection and prediction model based on explainable artificial intelligence for Alzheimer's disease. Sci. Rep. 11 (1), 1–26. doi:10.1038/s41598-021-82098-3.

Fabius, O., van Amersfoort, J.R., 2015. Variational recurrent auto-encoders. In: 3rd Int. Conf. Learn. Represent. ICLR 2015 - Work. Track Proc.. International Conference on Learning Representations, ICLR.

Fisher, C.K., Smith, A.M., Walsh, J.R., Simon, A.J., Edgar, C., Jack, C.R., Holtzman, D., Russell, D., Hill, D., Grosset, D., Wood, F., Vanderstichele, H., Morris, J., Blennow, K., Marek, K., Shaw, L.M., Albert, M., Weiner, M., Fox, N., Aisen, P., Cole, P.E., Petersen, R., Sherer, T., Kubick, W., 2019. Machine learning for comprehensive forecasting of Alzheimer's disease progression. Sci. Rep. 9 (1). doi:10.1038/s41598-019-49656-2. 1807.03876.

Fonteijn, H.M., Modat, M., Clarkson, M.J., Barnes, J., Lehmann, M., Hobbs, N.Z., Scahill, R.I., Tabrizi, S.J., Ourselin, S., Fox, N.C., Alexander, D.C., 2012. An event-based model for disease progression and its application in familial Alzheimer's disease and Huntington's disease. Neuroimage 60 (3), 1880–1889. doi:10.1016/j.neuroimage.2012.01.062.

Goldberg, Y., 2017. Neural network methods for natural language processing. Synth. Lect. Hum. Lang. Technol. 10 (1), 1–311. doi:10.2200/S00762ED1V01Y201703HLT037.

Gregor, K., Danihelka, I., Graves, A., Rezende, D.J., Wierstra, D., 2015. DRAW: a recurrent neural network for image generation. In: 32nd Int. Conf. Mach. Learn. ICML 2015. International Machine Learning Society (IMLS), pp. 1462–1471.

Hardoon, D. R., Szedmak, S., Shawe-Taylor, J., 2004. Canonical correlation analysis: an overview with application to learning methods. 10.1162/0899766042321814

Hochreiter, S., Schmidhuber, J., 1997. Long short-term memory. Neural Comput. 9 (8), 1735–1780. doi:10.1162/neco.1997.9.8.1735.

Hyun, J.W., Li, Y., Huang, C., Styner, M., Lin, W., Zhu, H., 2016. STGP: Spatio-temporal Gaussian process models for longitudinal neuroimaging data. Neuroimage 134, 550–562. doi:10.1016/j.neuroimage.2016.04.023.

Jack, C.R., Knopman, D.S., Jagust, W.J., Shaw, L.M., Aisen, P.S., Weiner, M.W., Petersen, R.C., Trojanowski, J.Q., 2010. Hypothetical model of dynamic biomarkers of the Alzheimer's pathological cascade. Lancet Neurol. 9 (1), 119–128. doi:10.1016/S1474-4422(09)70299-6.

Jacobs, H.I.L., Hopkins, D.A., Mayrhofer, H.C., Bruner, E., van Leeuwen, F.W., Raaijmakers, W., Schmahmann, J.D., 2017. The cerebellum in Alzheimer's disease: evaluating its role in cognitive decline. Brain 141 (1), 37–47. doi:10.1093/brain/awx194.

Kingma, D.P., Ba, J.L., 2015. Adam: a method for stochastic optimization. In: 3rd Int. Conf. Learn. Represent. ICLR 2015 - Conf. Track Proc.. International Conference on Learning Representations, ICLR.

Klami, A., Virtanen, S., Leppaaho, E., Kaski, S., 2015. Group factor analysis. IEEE Trans. Neural Netw. Learn. Syst. 26 (9), 2136–2147. doi:10.1109/TNNLS.2014.2376974. 1411.5799.

Lane, C.A., Hardy, J., Schott, J.M., 2018. Alzheimer's disease. Eur. J. Neurol. 25 (1), 59–70. doi:10.1111/ene.13439.

Édith Le Floch, Guillemot, V., Frouin, V., Pinel, P., Lalanne, C., Trinchera, L., Tenenhaus, A., Moreno, A., Zilbovicius, M., Bourgeron, T., Dehaene, S., Thirion, B., Poline, J.-B., Édouard Duchesnay, 2012. Significant correlation between a set of genetic polymorphisms and a functional brain network revealed by feature selection and sparse partial least squares. Neuroimage 63 (1), 11–24. doi:10.1016/j.neuroimage.2012.06.061.

Lee, G., Nho, K., Kang, B., Sohn, K.A., Kim, D., Weiner, M.W., Aisen, P., Petersen, R., Jack, C.R., et al., 2019. Predicting Alzheimer's disease progression using multi-modal deep learning approach. Sci. Rep. 9 (1), 1–12. doi:10.1038/s41598-018-37769-z.

Lei, B., Yang, M., Yang, P., Zhou, F., Hou, W., Zou, W., Li, X., Wang, T., Xiao, X., Wang, S., 2020. Deep and joint learning of longitudinal data for Alzheimer's disease prediction. Pattern Recognit. 102. doi:10.1016/j.patcog.2020.107247.

Lorenzi, M., Filippone, M., Frisoni, G.B., Alexander, D.C., Ourselin, S., 2019. Probabilistic disease progression modeling to characterize diagnostic uncertainty: application to staging and prediction in Alzheimer's disease. Neuroimage 190, 56–68. doi:10.1016/j.neuroimage.2017.08.059.

Lorenzi, M., Ziegler, G., Alexander, D.C., Ourselin, S., 2015. Efficient Gaussian process-based modelling and prediction of image time series. Inf. Process. Med. Imaging 24, 626–637.

Marinescu, R.V., Eshaghi, A., Alexander, D.C., Golland, P., 2019. BrainPainter: a software for the visualisation of brain structures, biomarkers and associated pathological pro-

cesses. In: Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics). Springer, pp. 112–120. doi:10.1007/978-3-030-33226-6_13.

Martí-Juan, G., Sanroma-Guell, G., Piella, G., 2020. A survey on machine and statistical learning for longitudinal analysis of neuroimaging data in Alzheimer's disease. Comput. Methods Programs Biomed. 189, 105348. doi:10.1016/j.cmpb.2020.105348.

Mehdipour Ghazi, M., Nielsen, M., Pai, A., Cardoso, M.J., Modat, M., Ourselin, S., Sørensen, L., 2019. Training recurrent neural networks robust to incomplete data: application to Alzheimer's disease progression modeling. Med. Image Anal. 53, 39–46. doi:10.1016/j.media.2019.01.004.

Molamohammadi, M., Rezaei-Shoshtari, S., Quitoriano, N., 2020. Jacobian of generative models for sensitivity analysis of photovoltaic device processes. Machine Learning for Engineering Workshop at NeurIPS, Vol. 2020.

Molchanov, D., Ashukha, A., Vetrov, D., 2017. Variational dropout sparsifies deep neural networks. In: 34th Int. Conf. Mach. Learn., ICML 2017, Vol. 5, pp. 3854–3863.

Moore, P.J., Lyons, T.J., Gallacher, J., 2019. Random forest prediction of Alzheimer's disease using pairwise selection from time series data. PLoS ONE 14 (2), e0211558. doi:10.1371/journal.pone.0211558. 1808.03273.

Mueller, S.G., Weiner, M.W., Thal, L.J., Petersen, R.C., Jack, C., Jagust, W., Trojanowski, J.Q., Toga, A.W., Beckett, L., 2005. The Alzheimer's disease neuroimaging initiative. Neuroimaging Clin. N. Am. 15 (4), 869–877. doi:10.1016/j.nic.2005.09.008.

Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., Ng, A.Y., 2011. Multimodal deep learning. In: Proc. 28th Int. Conf. Mach. Learn., ICML 2011, pp. 689–696.

Nguyen, M., Sun, N., Alexander, D.C., Feng, J., Thomas Yeo, B.T., 2018. Modeling Alzheimer's disease progression using deep recurrent neural networks. 2018 Int. Work. Pattern Recognit. Neuroimaging, PRNI 2018. Institute of Electrical and Electronics Engineers Inc. doi:10.1109/PRNI.2018.8423955.

Nie, L., Zhang, L., Meng, L., Song, X., Chang, X., Li, X., 2017. Modeling disease progression via multisource multitask learners: a case study with Alzheimer's disease. IEEE Trans. Neural Netw. Learn. Syst. 28 (7), 1508–1519. doi:10.1109/TNNLS.2016.2520964.

Oxtoby, N.P., Alexander, D.C., 2017. Imaging plus X: multimodal models of neurodegenerative disease. Curr. Opin. Neurol. 30 (4), 371–379. doi:10.1097/WCO.0000000000000460.

Reuter, M., Rosas, H.D., Fischl, B., 2010. Highly accurate inverse consistent registration: arobust approach. Neuroimage 53 (4), 1181–1196. doi:10.1016/j.neuroimage.2010.07.020.

Reuter, M., Schmansky, N.J., Rosas, H.D., Fischl, B., 2012. Within-subject template estimation for unbiased longitudinal image analysis. Neuroimage 61 (4), 1402–1418. doi:10.1016/j.neuroimage.2012.02.084.

Shaw, L.M., Vanderstichele, H., Knapik-Czajka, M., Clark, C.M., Aisen, P.S., Petersen, R.C., Blennow, K., Soares, H., Simon, A., Lewczuk, P., Dean, R., Siemers, E., Potter, W., Lee, V.M., Trojanowski, J.Q., 2009. Cerebrospinal fluid biomarker signature in Alzheimer's disease neuroimaging initiative subjects. Ann. Neurol. 65 (4), 403–413. doi:10.1002/ana.21610.

Shi, Y., Siddharth, N., Paige, B., Torr, P.H., 2019. Variational mixture-of-experts autoencoders for multi-modal deep generative models. NeurIPS.

Sobol', I., Kucherenko, S., 2009. Derivative based global sensitivity measures and their link with global sensitivity indices. Math. Comput. Simul. 79 (10), 3009–3017. doi:10.1016/j.matcom.2009.01.023.

Tsai, Y.H.H., Liang, P.P., Zadeh, A., Morency, L.P., Salakhutdinov, R., 2019. Learning factorized multimodal representations. In: Int. Conf. Learn. Represent. ICLR 2019 - Conf. Track Proc..

Verbeke, G., Fieuws, S., Molenberghs, G., Davidian, M., 2014. The analysis of multivariate longitudinal data: areview. Stat. Methods Med. Res. 23 (1), 42–49. doi:10.1177/0962280212445834.

Vounou, M., Nichols, T.E., Montana, G., 2010. Discovering genetic associations with high-dimensional neuroimaging phenotypes: a sparse reduced-rank regression approach. Neuroimage 53 (3), 1147–1159. doi:10.1016/j.neuroimage.2010.07.002.

Wang, T., Qiu, R.G., Yu, M., 2018. Predictive modeling of the progression of Alzheimer's disease with recurrent neural networks. Sci. Rep. 8 (1), 9161. doi:10.1038/s41598-018-27337-w.

Wu, M., Goodman, N., 2018. Multimodal generative models for scalable weakly-supervised learning. NeurIPS.

Yi, H.-A., Möller, C., Dieleman, N., Bouwman, F.H., Barkhof, F., Scheltens, P., van der Flier, W.M., Vrenken, H., 2016. Relation between subcortical grey matter atrophy and conversion from mild cognitive impairment to alzheimer's disease. J. Neurol. Neurosurg. Psychiatry 87 (4), 425–432. doi:10.1136/jnnp-2014-309105.

Young, A.L., Marinescu, R.V., Oxtoby, N.P., Bocchetta, M., Yong, K., Firth, N.C., Cash, D.M., Thomas, D.L., Dick, K.M., Cardoso, J., van Swieten, J., Borroni, B., Galimberti, D., Masellis, M., Tartaglia, M.C., Rowe, J.B., Graff, C., Tagliavini, F., Frisoni, G.B., Laforce, R., Finger, E., de Mendonça, A., Sorbi, S., Warren, J.D., Crutch, S., Fox, N.C., Ourselin, S., Schott, J.M., Rohrer, J.D., and, D.C.A., 2018. Uncovering the heterogeneity and temporal complexity of neurodegenerative diseases with subtype and stage inference. Nat. Commun. 9 (1). doi:10.1038/s41467-018-05892-0.

Young, A.L., Oxtoby, N.P., Daga, P., Cash, D.M., Fox, N.C., Ourselin, S., Schott, J.M., Alexander, D.C., 2014. A data-driven model of biomarker changes in sporadic Alzheimer's disease. Brain 137 (Pt 9), 2564–2577. doi:10.1093/brain/awu176.

Young, A.L., Oxtoby, N.P., Huang, J., Marinescu, R.V., Daga, P., Cash, D.M., Fox, N.C., Ourselin, S., Schott, J.M., Alexander, D.C., 2015. Multiple orderings of events in disease progression. In: Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics). Springer, Cham, pp. 711–722. doi:10.1007/978-3-319-19992-4_56.

Young, A.L., Vogel, J.W., Aksman, L.M., Wijeratne, P.A., Eshaghi, A., Oxtoby, N.P., Williams, S.C., Alexander, D.C., 2021. Ordinal SuStaIn: subtype and stage inference for clinical scores, visual ratings, and other ordinal data. Front. Artif. Intell. 4 (August), 1–13. doi:10.3389/frai.2021.613261.