Multi-task longitudinal forecasting with missing values on Alzheimer's Disease

Carlos Sevilla-Salcedo[a,*], Vandad Imani[b], Pablo M. Olmos[a], Vanessa Gómez-Verdejo[a], Jussi Tohka[b], for the Alzheimer's Disease Neuroimaging Initiative[1]

[a] *Signal Theory and Communications Department, University Carlos III of Madrid, Leganés, 28911 Spain*
[b] *A.I. Virtanen Institute for Molecular Sciences, University of Eastern Finland, Kuopio, Finland*

Abstract

Background and Objective: Machine learning techniques typically used in dementia assessment are not able to learn multiple tasks jointly and deal with time-dependent heterogeneous data containing missing values. In this paper, we reformulate SSHIBA, a recently introduced Bayesian multi-view latent variable model, for jointly learning diagnosis, ventricle volume, and ADAS score in dementia on longitudinal data with missing values.

Methods: We propose a novel Bayesian Variational inference framework capable of simultaneously imputing missing values and combining information from several views. This way, we can combine different data views from different time-points in a common latent space and learn the relationships

[*] Corresponding author.
Email address: `sevisal@tsc.uc3m.es` (Carlos Sevilla-Salcedo)

[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. This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Institutional Data Access/Ethics Committee of the ADNI. A complete listing of ADNI investigators can be found at: `http://adni.loni.usc.edu/wpcontent/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf`. This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Institutional Data Access/Ethics Committee of the Alzheimer's Disease Neuroimaging Initiative (ADNI).

between each time-point, using the semi-supervised formulation to fully exploit the temporal structure of the data and handle missing values. In turn, the model can combine all the available information to simultaneously model and predict multiple output variables.

Results: We applied the proposed model to jointly predict diagnosis, ventricle volume, and ADAS score in dementia. The comparison of imputation strategies demonstrated the superior performance of the semi-supervised formulation of the model, improving the best baseline methods. Moreover, the performance in simultaneous prediction of diagnosis, ventricle volume, and ADAS score led to an improved prediction performance over the best baseline method.

Conclusions: The results demonstrate that the proposed SSHIBA framework can learn an excellent imputation of the missing values and outperforming the baselines while simultaneously predicting three different tasks.

Keywords: Alzheimer's disease, longitudinal data, missing values, multi-task

1. Introduction

Alzheimer's Disease (AD) is a common form of dementia that manifests in the form of cognitive degeneration and conduct disorder. Although its symptoms vary between subjects, it is commonly characterised by memory loss as well as general cognitive decline. More than 30 million people suffer from AD currently and this number is expected to triple by 2050 [1]. The number of people affected by the disease is higher than the number of AD patients due to the huge impact on the lives of relatives, friends and care-givers. AD has no cure, but interventions taking place early on during the disease cascade can improve the quality of life and alleviate the symptoms [2]. For this reason, the investigations for an early detection of AD in high risk individuals are critical. Similarly, cognitive scores are essential for understanding the efficacy of antidementia treatments as well as the disease progression[3]. Machine Learning (ML) techniques can be used for the design of imaging biomarkers for various brain disorders and, additionally, the inferred ML models can be analysed as multivariate, discriminative representations of the brain disease.

Analysis of the progression of dementia in longitudinal studies has been proven to be critical for an adequate treatment[4]. Some algorithms primarily use neuroimaging information to analyse the disease progression of the disease[5]. Koval et al.[6] used graph nodes to represent a spatially structured mixed-effect model with a Monte-Carlo Markov-Chain Stochastic Approximation Expectation-Maximization to model the evolution of longitudinal brain imaging data. In contrast, other studies focus on analysing the progression of biomarkers to characterise the disease. Venkatraghavan et al.[7] construct a timeline of biomarker changes and models the brain abnormality with APOE genotypes using Gaussian Mixture Models to then calculate the probability of abnormality. Similarly, Donohue et al.[8] assume that the biomarkers are a set of curves with a common shape combined with simple linear effects at the subject level, while modelling long-term traits with nonparametric monotonic smoothing. However, some studies combine both neuroimaging and biomarker information in a single framework to model the longitudinal effect of AD. Fonteijn et al.[9] combines heterogeneous data using both imaging and clinical data by defining the disease as a sequence of discrete events using a Bayesian model.

Longitudinal dementia studies faces issues of missing data due to old age and health related concerns in the studied population[10]. This problem is greatly accentuated when working with longitudinal data, where follow up measures are often interrupted either by cognitive impairment or mortality[11]. To work around the issue, some studies remove all the data from subjects with missing values. However, this reduces the number of usable data samples and can introduce bias in the data[12]. For this reason, other studies[13] apply diverse techniques to impute the missing data. Some use basic inference techniques such as substituting the missing variable by its mean, median or mode value. Others exploit the longitudinal nature of the problem to infer the missing values using temporal imputation, using available information at previous months[14]. Adhikari et al.[15] combines both imputation techniques, using temporal inference if previous data is available and the variable median otherwise. Bayesian algorithms assume variables are random and learn their distribution, which can be used to impute these missing values, for instance, using the mean of the distribution or sampling from the distribution[16].

Multitask learning (MTL) is a sub-field of ML that simultaneously learns multiple tasks by jointly optimising multiple loss functions. MTL models improve generalisation by leveraging the information contained in the training data of related tasks. It is, therefore, beneficial when the tasks have some level of correlation. In recent years, MTL has attracted a lot of attention in the prediction of the progression of cognitive decline in dementia at multiple time points with clinical data [17, 18]. The fundamental hypothesis in the longitudinal analysis is that the subject's clinical data cannot be assumed to be independent at consecutive visits. Accordingly, the MTL can benefit the prediction of disease progression by capturing relatedness and shared information between several observation records across the visits. One of the critical issues in MTL is identifying the inherent relation between these records of observation and tasks. MTL methods with sparsity-inducing norm regularisation have been widely studied to improve generalisation performance by simultaneously solving multiple learning tasks while utilising commonalities across tasks. For example, Jin et al. [19] developed a feature level-based group lasso formulation as the regularization term to capture intrinsic relationships among tasks as well as supplementary information from other unrelated tasks at the feature level to classify patients with Mild Cognitive Impairment (MCI) and Normal Control (NC). Cao et al. [20] proposed a $\ell_{2,1}$ -ℓ_1-norm regularised multi-kernel MTL feature learning formulation with a joint sparsity inducing regularisation (SMKMTL). Their framework uses a mixed sparsity-Inducing $\ell_{2,1}$ -ℓ_1-norm to capture the inherent correlation among the tasks. They have proposed SMKMTL multitask learning method to capture the kernel-wise association between MRI neuroimaging features and cognitive scores. Yang et al. [21] proposed a fused sparse network algorithm with parameter-free centralised learning to model and identify the longitudinal analysis of early MCI and late MCI based on resting-state functional MRI. Tabarestani et al. [22] proposed a multitask multimodal framework for predicting cognitive measures in the progression of AD. They applied $\ell 1$-norm regularisation to introduce sparsity among all features and capture different modalities' inherent temporal sparsity patterns and their relative correlation strength.

However, we are not aware of studies or methods that analyse cognitive decline and combine MTL, longitudinal data and missing data imputation within a single framework. To address this, we propose developing a model based on the recently presented SSHIBA framework [23] to predict various facets of cognitive decline. This model has the ability to work with multiple views and impute missing values. Specifically, we propose establishing different time-points as different views in order to model temporal relations and make a forecast for future time-points. This allows the model to find a common latent representation of time-dependent and time-independent variables that describes the temporal relation between variables over time. Furthermore, the multi-view formulation of the framework allows having regression MTL. SSHIBA eliminates the need to train a model for each task and enhances the performance by combining the information of the multiple tasks in the common framework. Finally, SSHIBA allows to impute all missing values in any view using its semi-supervised formulation to learn the joint variable distribution of the train and test datasets. This allows to automatically impute missing values using the learnt distribution and, simultaneously, using a semi-supervised scheme to predict the test data.

To analyse the performance of SSHIBA, we use the Alzheimer's Disease Neuroimaging Initiative (ADNI) database. Based on previous research, we select a subset of relevant variables for characterising AD to analyse the time evolution of the subject's diagnosis, ADAS-cog13 score (ADAS13), and ventricle volume. We measure the performance against several baseline methods for the imputation of missing values and the multitask prediction of the output variables. The results indicate that SSHIBA outperforms the baselines on the prediction task while finding a latent representation of the data that improves the interpretability of the predictions.

2. Material and Methods

2.1. Material

Data used in the preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (`adni.loni.usc.edu`). The ADNI project was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the longitudinal progression of Mild Cognitive Impairment (MCI) and early AD. In particular, we use the tables prepared for the TADPOLE grand challenge based on ADNI data (`https://tadpole.grand-challenge`) [24]. Although in the original database there is information for up to 120 months after baseline, for this article, we use the data corresponding to the first 36 months.

From the original database with $1,739$ subjects we selected the $1,730$ subjects with any information on month 36, described in Table 1. Following [25], we use the features that are specially relevant for the characterisation of Alzheimer's disease (see Table 2 for a summary of these features). Figure 1 depicts the number of missing values for each Time Dependent (TD) variable at each analysed month. Note that the proportion of missing values is large for most features. This is because, although we use all variables at each visit (timepoint), not all variables are acquired at every visit and not all participants are scheduled to visit at every 6 months.

Diagnosis	No. of Subjects	Age	Male / Female
NC	523	74.22 ± 5.80	253 / 270
MCI	866	73.05 ± 7.60	512 / 354
AD	341	74.98 ± 7.76	189 / 152

Table 1: Demographic information of participants. The diagnosis is the diagnosis at the baseline. The column age lists the average age followed by the standard deviation of age.

We considered various participant details that have been found to be risk factors contributing to AD[26]: age, sex, the number of APOE e4 alleles, and the years of education. We coded APOE e4 status as either absence (0), single copy (1) or homozygous (2).

As FDG-PET features, we used average standardised uptake values (SUVs) in five brain regions from the ADNI database: bilateral angular gyri, bilateral posterior cingulate gyri, and bilateral inferior temporal gyri. The FDG-PET data measures glucose consumption and is shown to be strongly related to dementia and cognitive impairment when compared to normal control subjects[27, 28]. Motion correction and co-registration with MRI was performed on the acquired PET data. The highest 50% of voxel values within a hand-drawn pons/cerebellar vermis region were selected and their mean was used to normalise each ROI measurement resulting in the final FDG-PET measurements. As specified in [29, 30], FDG-PET has been criticized as longitudinal biomarker for the analysis of cognitive decline, whereas AV-45 PET and MRI are more powerful in the longitudinal analysis of the disease. For this reason, we include only the FDG-PET values at baseline for these experiments.

The neuropsychology and behavioral (NePB) assessments reflect the cognitive abilities of the subjects. Subjects underwent a battery of NePB tests[31]. We included 5 NePB scores as features: the summary score from Mini-Mental State Examination (MMSE)[32], three summary scores of Rey's auditory verbal learning test (RAVLT; learning, immediate, and percent forgetting)[33], and a summary score from the functional activities questionnaire (FAQ)[34].

As AV-45 PET features, we used SUVs in seven regions of interest (ROIs): frontal cortex, cingulate, lateral parietal cortex, lateral temporal cortex, cerebellum grey matter, whole cerebellum, and eroded subcortical white matter. The AV-45 PET measures amyloid-beta load in the brain. AV-45 PET imaging and preprocessing details are available at `http://adni.loni.usc.edu/methods/pet-analysis-method/pet-analysis/`. We used

Variable	Description	Group
Age	Subject details	TI
Sex		TI
APOE4		TI
Education		TI
AngularLeft	FDG-PET	TI
AngularRight		TI
CingulumPostBilateral		TI
TemporalLeft		TI
TemporalRight		TI
MMSE	NePB	TD
RAVLT learning		TD
RAVLT immediate		TD
RAVLT perc forgetting		TD
FAQ		TD
Cerebellum Grey Matter	AVF45 data	TD
Whole Cerebellum		TD
Eroded Subcortical Wm		TD
Frontal		TD
Cingulate		TD
Parietal		TD
Temporal		TD
ABETA	CSF values	TD
TAU		TD
PTAU		TD
Hippocampus	MRI volumetry	TD
WholeBrain		TD
Entorhinal		TD
Fusiform		TD
MidTemp		TD
ICV		TD
Ventricle volume	MRI volumetry	V
ADAS13	ADAS-Cog13 score	A
Diagnosis	Clinical diagnosis	D

Table 2: Description of the variables used in this study. Each variable is assigned to one group to facilitate the understanding of the framework: *TI*, time-independent variables; *TD*, time-dependent variables; *D*, Diagnosis; *V*, Ventricle volume; *A*, ADAS13 score. NePB indicates neuropsychology and behavioral tests and we use FDG-PET at baseline.

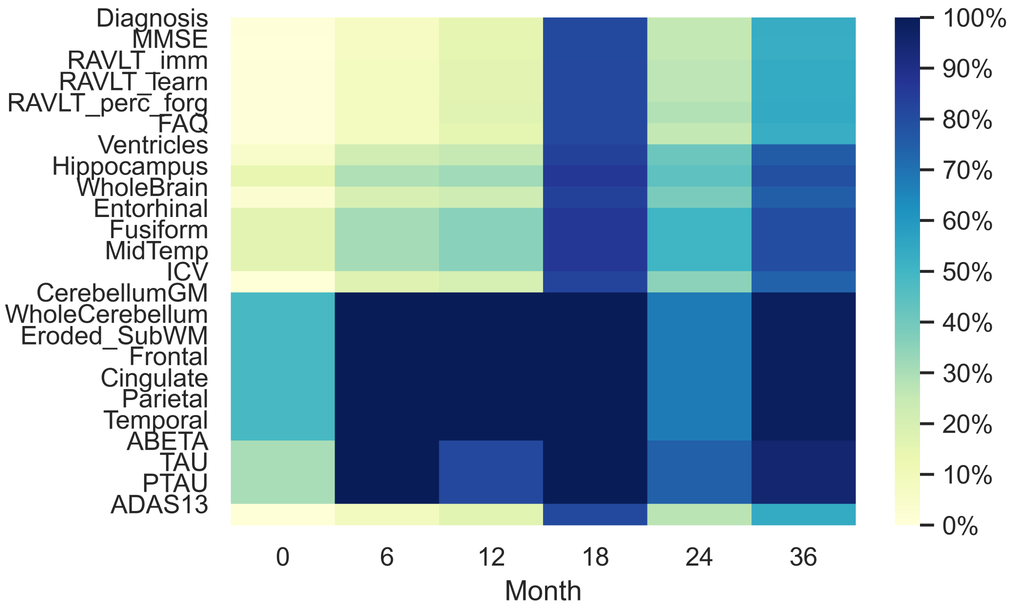


Figure 1: Heat map of the percentage of missing values for each TD variable for each month. Dark blue colours represent the variable is missing for all individuals at the specified month, while light yellow colours represent there are no missing values.

regional SUV ratios processed according to the UC Berkeley protocol[35, 36, 27]. Each AV-45 PET scan was co-registered to the corresponding MRI and the mean AV-45 uptake within the regions of interest and reference regions was calculated. We included the values in ROIs as well as values in the reference regions as variables to the model to learn to normalise the values as it remains uncertain which reference region would be the best for standardisation [37].

The baseline Cerebrospinal Fluid (CSF) Aβ42, t-tau, and p-tau were used as CSF features[38].

As MRI features, we used 7 features: intracranial volume (ICV), and volumes of the hippocampus, entorhinal cortex, fusiform gyri, whole brain, middle temporal gyri and lateral ventricles. These features were selected based on previous studies[39]. MRI protocol details are provided by ADNI at `http://adni.loni.usc.edu/methods/mri-tool/mri-analysis/`. Cortical reconstruction and volumetric segmentation had been performed with the FreeSurfer 5.1 image analysis suite.

The ADAS-cog 11 task scale was developed to assess the efficacy of anti-dementia treatments. Further developments to the scale shifted its sensitivity towards pre-dementia syndromes as well, primarily mild cognitive impairment (MCI). The ADAS-cog 13 task scale was one such improvement on the original ADAS-cog 11, with additional memory and attention/executive function tasks[40]. The 13 tasks test verbal memory (3 tasks), clinician-rated perception (4 tasks), and general cognition (6 tasks). It was found to perform better than the ADAS-cog 11 at discriminating between MCI and mild AD subjects, as well as have better sensitivity to treatment effects in MCI[41]. As such, we used the ADAS-cog 13 scale for our study as a continuous quantitative measure of a subject's disease status. The value of these scores is lowest for the normal control group and increases with disease progression, with the highest scores for AD subjects.

As in the TADPOLE competition [24, 42], we consider: the clinical diagnosis (NC, MCI, AD) denoted as D, the ventricle volume of the MRI data denoted as V and the ADAS-cog 13 score denoted as A.

2.2. SSHIBA framework

In this work, we adapt the recently presented SSHIBA framework[23] to model longitudinal data. SSHIBA is a Bayesian model able to combine multiview heterogeneous information into a common latent space. Here, we propose exploiting this multiview formulation to model the progression of the variables over time.

We consider a multi-view problem where we have N data samples represented in M different views, $\{\mathbf{X}^{(\mathrm{m})}\}_{m=1}^{M}$. Therefore, we have that $\mathbf{x}_{\mathrm{n},:}^{(\mathrm{m})} \in \mathbb{R}^{1\times D_m}$ is the m-th view of the n-th subject with $n = 1, \dots, N$ and $\mathcal{M} = \{1, \dots, M\}$, so $\mathbf{x}_{\mathrm{n},:}^{\{\mathcal{M}\}} = \{\mathbf{x}_{\mathrm{n},:}^{(1)}, \dots, \mathbf{x}_{\mathrm{n},:}^{(\mathrm{M})}\}$ is the complete n-th observation[1]. The model considers there is a continuous latent variable, $\mathbf{Z}$, that can be combined with a set of projection matrices for each view, $\mathbf{W}^{\{\mathcal{M}\}} = \{\mathbf{W}^{(1)}, \dots, \mathbf{W}^{(\mathrm{M})}\}$, and add some gaussian noise, with precision $\tau^{\{\mathcal{M}\}} = \{\tau^{(1)}, \dots, \tau^{(\mathrm{M})}\}$, to generate the complete observation $\mathbf{x}_{\mathrm{n},:}^{\{\mathcal{M}\}}$. Then, the probability density function

[1] Each view comprises an observation and the set of random variables associated in the model.

(pdf) of the random variables for each view can be defined as

$$\mathbf{z}_{\mathrm{n},:} \sim \mathcal{N}(0, I_{K_c}) \tag{1}$$

$$\mathbf{w}^{(\mathrm{m})}_{:,\mathrm{k}} \sim \mathcal{N}\left(0, \left(\alpha^{(\mathrm{m})}_{\mathrm{k}}\right)^{-1} I_{K_c}\right) \tag{2}$$

$$\mathbf{x}^{(\mathrm{m})}_{\mathrm{n},:} \,|\, \mathbf{z}_{\mathrm{n},:} \sim \mathcal{N}(\mathbf{z}_{\mathrm{n},:}\, \mathbf{W}^{(\mathrm{m})^{\mathrm{T}}}, \tau^{(\mathrm{m})^{-1}}\, I_{D_m}) \tag{3}$$

$$\alpha^{(m)}_{k} \sim \Gamma\left(a^{\alpha^{(m)}}, b^{\alpha^{(m)}}\right) \tag{4}$$

$$\tau^{(\mathrm{m})} \sim \Gamma\left(a^{\tau^{(\mathrm{m})}}, b^{\tau^{(\mathrm{m})}}\right) \tag{5}$$

where I_{K_c} is an identity matrix of dimension K_c, $\mathbf{z}_{\mathrm{n},:} \in \mathbb{R}^{1\times K_c}$ is the low-dimension latent variable for the n-th data point, $\Gamma(a, b)$ is a Gamma distribution with parameters a and b, $\mathbf{w}^{(\mathrm{m})}_{:,\mathrm{k}}$ is the k-th column of matrix $\mathbf{W}^{(\mathrm{m})}$ (of dimensions $D_m \times K_c$), and up-script (m) corresponds to the m-th view. While the Gaussian distribution adjusts to the continuous nature of the projection variables, the Gamma distribution over $\alpha^{(m)}_{k}$ enables the model to enforce zero values in order to maximise the model likelihood given our data. Hence, we say that (2) and (4) form an ARD (Automatic Relevance Determination) prior[43] for each column of matrix $\mathbf{W}^{(\mathrm{m})}$. The SSHIBA graphical model for the generation of each data view is included in Figure 2.

These equations correspond to the standard formulation of SSHIBA to model real observations. However, if besides modelling real observations, we require using feature selection in a certain view, we can modify Equation (2) and add a new variable $\boldsymbol{\gamma}^{(\mathrm{m})}$

$$\mathrm{w}^{(\mathrm{m})}_{\mathrm{d,k}} \sim \mathcal{N}\left(0, \left(\gamma^{(\mathrm{m})}_{\mathrm{d}}\, \alpha^{(\mathrm{m})}_{\mathrm{k}}\right)^{-1}\right) \tag{6}$$

$$\gamma^{(\mathrm{m})}_{\mathrm{d}} \sim \Gamma\left(a^{\boldsymbol{\gamma}^{(\mathrm{m})}}, b^{\boldsymbol{\gamma}^{(\mathrm{m})}}\right) \tag{7}$$

where $\gamma^{(\mathrm{m})}_{\mathrm{d}}$ is equivalent to $\alpha^{(\mathrm{m})}_{\mathrm{k}}$, having that (6), (4) and (7) form the ARD prior for each row and column of matrix $\mathbf{W}^{(\mathrm{m})}$. This way, while $\alpha^{(\mathrm{m})}_{\mathrm{k}}$ forces zero values column-wise in the latent variables, $\gamma^{(\mathrm{m})}_{\mathrm{d}}$ forces the zero values in the features, having a double automatic selection of relevant features.

After defining the generative model, we can evaluate the posterior distribution of all the model variables using an approximate inference approach through mean-field variational inference [44]. With this, we maximise a lower

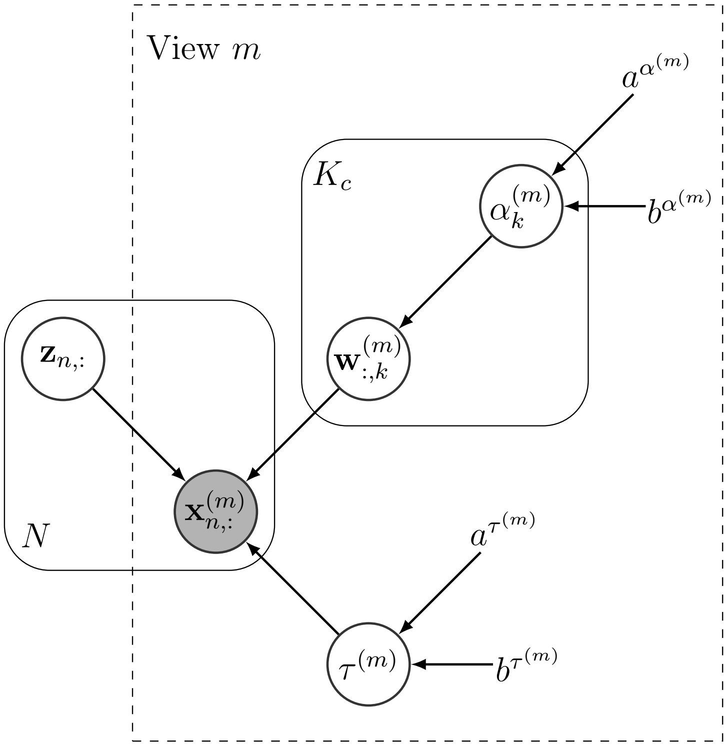


Figure 2: Plate diagram for the SSHIBA graphical model with one view and real valued observations. Gray circles denote observed variables, white circles unobserved random variables. The nodes without a circle correspond to the hyperparameters.

bound to the posterior distribution and choose a fully factorised variational family to approximate the posterior distribution as

$$p(\Theta|\,\mathbf{X}^{\{\mathcal{M}\}}) \approx \prod_{\mathrm{n=1}}^{\mathrm{N}} q(\mathbf{z}_{\mathrm{n,:}})$$
$$\prod_{\mathrm{m=1}}^{\mathrm{M}} \left(q\big(\mathbf{W}^{(\mathrm{m})}\big) q\big(\tau^{(\mathrm{m})}\big) \prod_{\mathrm{k=1}}^{\mathrm{K_c}} q\Big(\alpha_{\mathrm{k}}^{(\mathrm{m})}\Big) \prod_{\mathrm{d=1}}^{\mathrm{D_m}} q\Big(\gamma_{\mathrm{d}}^{(\mathrm{m})}\Big) \right), \tag{8}$$

where Θ comprises all random variables in the model and $\mathcal{M}_i$ represents the set of views with binary data.

The mean-field posterior structure along with the lower bound results in a feasible coordinate-ascent-like optimisation algorithm in which the optimal maximisation of each of the factors in (8) can be computed if the rest remain fixed using the following expression

$$q^*(\theta_i) \propto \mathbb{E}_{\Theta_{-i}} \left[\log p(\Theta, \mathbf{x}_{1,:}, \ldots, \mathbf{x}_{N,:})\right], \tag{9}$$

where Θ_{-i} comprises all random variables but θ_i. This new formulation is in general feasible since it does not require to completely marginalise Θ from the joint distribution.

2.3. SSHIBA implementation on longitudinal data

Taking advantage of SSHIBA formulation, we propose utilising the multi-view framework to combine time-independent and time-dependent variables (as specified in Table 2). To do so, we firstly defined one view in charge of modelling the time-independent observations, $\mathbf{X}^{(1)}$. Then, we combined the time-dependent data in various views based on their time-stamps. This way, we have the measures corresponding to time-stamps (6, 12, 18, 24 and 30 months before the prediction) modelled each in one view, $\{\mathbf{X}^{(2)}, \ldots, \mathbf{X}^{(6)}\}$. Finally, we have one view for each prediction task $\{\mathbf{X}^{(12)}, \mathbf{X}^{(13)}, \mathbf{X}^{(14)}\}$ at the desired month. Figure 3 summarises the defined views where we also included views $7, \ldots, 12$ to model the diagnosis as a one-vs-all observation. Therefore, we use the information of months 0, 6, 12, 18 and 24 to predict a result 30 months after the baseline[2].

This way, the model is capable of learning a projection matrix $\mathbf{W}^{(\mathrm{m})}$ for each time-stamp and a latent variable $\mathbf{Z}$ in charge of defining the relation between the views through the latent space. This allows the model to learn common factors corresponding to temporal information, but also to learn timestamp specific latent variables. Besides, this information can be combined with time-independent variables while analysing through this latent representation the relation with the output tasks. Therefore, in this framework, we use information of months $0, \ldots, 24$ to predict the output variables at month 30 to train the model. For testing we use months $6, \ldots, 24$ to predict the output variables at month 36, where we do not include information of month 30 to have year prediction.

In order to improve the model performance, we propose using the SSHIBA's missing values imputation functionality to increase the number of samples per subject used to train the model by including missing values in the months

[2] Note that we are doing a variable forecasting of time-stamp [t], therefore, we do not use any information of that time-stamp for the prediction task.

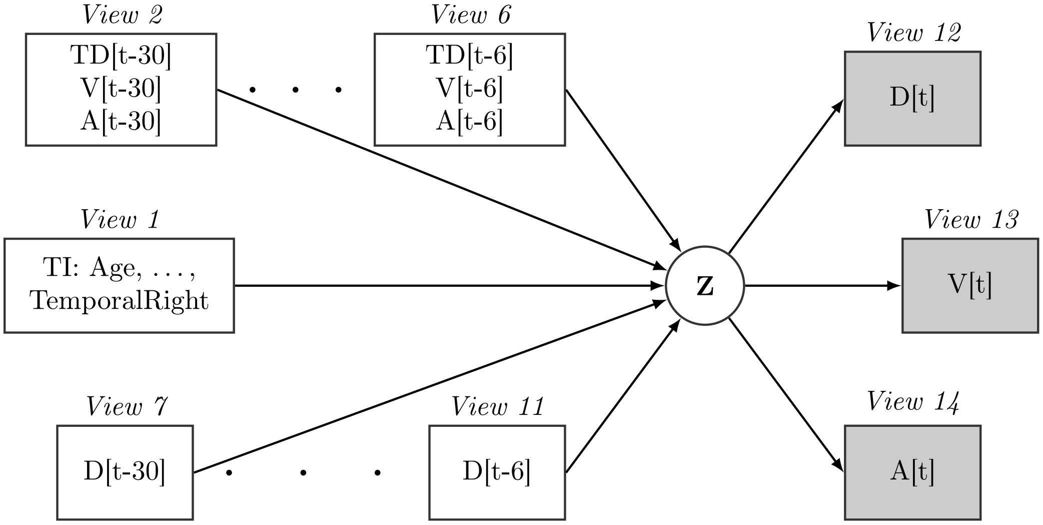


Figure 3: Plate diagram for the multi-view SSHIBA model adapted to longitudinal data. White circles denote unobserved random variables, white squares denote input views and grey squares output views. Note that each view is comprised by its observations and its corresponding random variables. *TI* comprises the time-independent variables and *TD[t]* the time-dependent variables without the *Diagnosis* for month t (*D[t]*), the ventricle volume for month t (*V[t]*) and the ADAS13 score for month t (*A[t]*).[t-6], . . . , [t-30] represent 6,. . . , 30 months before month t.

previous to the baseline. Therefore, we do not just use the information related to months 0, 6, 12, 18 and 24 to predict 30 as our training data, but also months -6, 0, 6, 12 and 18 to predict 24 and so on, where every negative month is filled with missing values. This can be done thanks to the shared latent representation, **Z**, which establishes the relation between each time-stamp and, in turn, uses the information of available information from different time-stamps to impute missing values. This scheme is summarised in Table 3, having that the test set is always set to predict month 36 and always uses months 6 to 24 as input. This implies that the model is calculated without using any information from month 36. This change in the way we handle the information allows the model to have data augmentation with 5 times more training samples, $8,685$ training samples, as well as using all the available data to train the model. Table 3 includes the structure for the train and test set used in this article.

During the inference learning we remove the k columns of $\mathbf{W}^{(\mathrm{m})}$, $\forall m$, if all the elements of $\mathbf{w}_{:,\mathrm{k}}^{(\mathrm{m})}$, across all views, are lower than the pruning threshold.

	View	Samples for subject n					
		train 1	train 2	train 3	train 4	train 5	test 1
Input	1	*TI*	*TI*	*TI*	*TI*	*TI*	*TI*
	2	–	–	–	–	*TD[0]*	*TD[6]*
	3	–	–	–	*TD[0]*	*TD[6]*	*TD[12]*
	4	–	–	*TD[0]*	*TD[6]*	*TD[12]*	*TD[18]*
	5	–	*TD[0]*	*TD[6]*	*TD[12]*	*TD[18]*	*TD[24]*
	6	*TD[0]*	*TD[6]*	*TD[12]*	*TD[18]*	*TD[24]*	–
	7	–	–	–	–	*D[0]*	*D[6]*
	8	–	–	–	*D[0]*	*D[6]*	*D[12]*
	9	–	–	*D[0]*	*D[6]*	*D[12]*	*D[18]*
	10	–	*D[0]*	*D[6]*	*D[12]*	*D[18]*	*D[24]*
	11	*D[0]*	*D[6]*	*D[12]*	*D[18]*	*D[24]*	–
Output	12	*D[6]*	*D[12]*	*D[18]*	*D[24]*	–	*D[36]*
	13	*V[6]*	*V[12]*	*V[18]*	*V[24]*	–	*V[36]*
	14	*A[6]*	*A[12]*	*A[18]*	*A[24]*	–	*A[36]*

Table 3: Data configuration for the framework. *TI* represents the time-independent data, *D* represents the diagnosis, *V* the ventricle volume, *A* the ADAS13 score and *TD* the time-dependent data for month t including *V* and *A*. A hyphen implies that we are not using any information on that view for that sample and we are just using missing values. The training results corresponding to month 30 are set to missing values in order to do a one year prediction of the test set.

For our experiments, this pruning threshold was set to 10^{-6}. To determine the number of iterations of the inference process, we used a convergence criteria based on the evolution of the lower bound. In particular, we stop the algorithm either when $LB[-2] > LB[-1](1-10^{-6})$, where $LB[-1]$ is the lower bound at the last iteration and $LB[-2]$ at the previous one, or when it reaches 5×10^4 iterations. We used learning rates of $1/\#iter$, 1 and 0.9 for the projection matrices of views 7-12, 14 and the rest to adapt to the MTL problem. The SSHIBA model is available at `https://github.com/sevisal/SSHIBA.git`.

3. Results

This section presents the results of the prediction of clinical diagnosis, ventricle volume and the ADAS13. We additionally compare the SSHIBA based model to selected state-of-the-art methods as well as standard baseline methods.

3.1. Baseline methods

We used Ridge Regression (RR) and Logistic Regression (LogR) as baseline regression and classification algorithms to analyse the imputation performance on A and V and D, respectively. We also selected several multitask learning based methods for comparison. First, we included a convex fused sparse group Lasso[13] (cFSGL) formulation. This technique encodes the temporal information by considering the sparse group Lasso penalty to select a common set of biomarkers across multiple time points and simultaneously incorporate temporal smoothness using the fused Lasso penalty. Second, we also considered multi-task techniques with concatenated temporal information:

- ℓ_1-norm regularized multitask learning (Least Lasso)[45]: The $\ell 1$-norm regularisation term captures the task relationship from multiple related tasks which introduce sparsity into the features along with the parameter for controlling the sparsity among all tasks.

- Joint Feature Selection (JFS)[46]: it utilises the $\ell_{1,2}$-norm regularisation term to learn sparse representations across multiple related tasks to constrain all models to share a common set of features.

- Dirty Model[47]: it explicitly estimates a sum of two sets of parameters with multiple individual regularisation, where the corresponding matrices are encouraged to have element-wise sparsity and block-structured row sparsity.

- Low Rank Assumption (LRA)[48]: it captures the task relationships using a low-rank structure, and simultaneously identifies outlier tasks using a group-sparse structure.

The MALSAR package[49] running in MATLAB was used to implement the MTL algorithms. The regularisation parameters ρ_1 (the regularisation parameter controlling the sparsity among all tasks) and ρ_2 (an optional regularisation parameter that controls the $\ell 2$-norm penalty) are selected by 10-fold cross-validation strategy on the training data. The ρ_1 and ρ_2 parameters were selected among the candidate set $\{10^{-3}, 10^{-2.5}, \ldots, 10^{2}, 2 \times 10^{2}, 2.5 \times 10^{2}, \ldots, 5 \times 10^{2}\}$ by minimising the mean absolute error (MAE) for the regression tasks and balanced accuracy for the classification tasks. The regularisation parameter α for RR was selected using 10-fold cross-validation from a grid of 11 values that were logarithmically spaced between -20 and 2.

Finally, we also included two baselines that rely on latent representations using Bayesian formulation:

1. Manifold Relevance Determination (MRD) [50]. This model relies on a shared GPLVMs approach and includes an ARD for relevance vectors selection. We used the available library in *Matlab*, setting the number of latents to the number of tasks ($K = C$) and using 500 relevance vectors.

2. Heterogeneous Incomplete - Variational AutoEncoder (HI-VAE) [51]. This model is also capable of working with missing values while finding low-dimensional data representations. We decided to use the layer configuration suggested by the authors, two layers of dimensions 50-50-50.

For these methods we also considered a one-vs-all codification of the diagnosis. Therefore, they will carry out 5 task learning problem. We included all the input views as input variables for each model.

To impute the missing data for all baselines we used 5 distinct strategies: substituting by zero, the mean, the median and the most frequent value and temporal imputation. Temporal imputation of the missing values consists in substituting them by the mean of the subject's previous existing values if available and by the mean of the variable otherwise.

3.2. Metrics

For the regression tasks, prediction of ADAS13 and Ventricle volume at month 36 (views 13 and 14), we quantified the predictive accuracy using the Mean Absolute Error (MAE) calculated as $\mathrm{MAE} = \frac{1}{N}\sum_{\mathrm{n=1}}^{\mathrm{N}} \left|\mathbf{y}_n^{true} - \mathbf{y}_n^{pred}\right|$, where $\mathbf{y}_n^{true}$ is the true value for sample n and $\mathbf{y}_n^{pred}$ is the predicted value for sample n. We also included a hypothesis test over these tasks using the difference between the error mean of SSHIBA and the error mean of the corresponding baseline as a test statistic. For the classification task, prediction of variable Diagnosis at month 36 (view 12), we quantified the predictive accuracy using the balanced multiclass Area Under the Curve (mAUC) [52] calculated as $\mathrm{mAUC} = \frac{1}{N}\sum_c(N_c \times AUC_c)$, where N_c is the number of samples of class c and AUC_c is the AUC of class c with respect to the rest of the classes.

3.3. Performance compared to baselines

In this section, we analyse the scores obtained in the prediction of the three output variables. First we compare imputation strategies in a single variable prediction problem, where we only predict variable A (The equivalent results on V and D are available in the supplementary material). Table 4 depicts the performance obtained by RR and SSHIBA where SSHIBA automatically imputes the missing values and RR uses distinct imputation techniques. Besides, we calculate the MAE on the prediction of ADAS13 at month 36 using various data information for the months previous to 36. Specifically, A implies that we only use ADAS13 information on previous time-stamps to predict the value at month 36, Multimodal Data (MD) represents the variables that are not ADAS13 and MD + A that we use all the available variables. The results determine that using temporal imputation improves the performance of RR. However, SSHIBA greatly outperforms any of the reference baselines independently of the input variables we use to train the model. Note that SSHIBA improves its performance combining all the available information whereas RR+temporal hinders it performance when using the rest of the data.

Regressor	Imputation	Input features		
	strategy	A	MD	MD + A
RR	*zero*	11.201	10.907	7.765
	mean	5.766	5.332	5.194
	median	5.957	5.657	5.234
	most frequent	8.220	8.095	6.257
	temporal	4.045	4.495	4.258
SSHIBA		3.613	4.012	**3.407**

Table 4: Results obtained in the prediction of ADAS13 score at month 36 using information from baseline to month 24. We used MAE score as a performance measure. Columns A, MD and MD+A show the results obtained using only ADAS13 score, MD and both as input, respectively.

In a second experiment we compare the performance of the baselines with SSHIBA where we simultaneously predict the three output variables.

Based on the previous results, we impute the baseline missing values using the temporal imputation.

Model	A	V	D
Least Lasso	3.623^{*}	$3,981^{***}$	0.953
JFS	3.760^{**}	$3,952^{***}$	0.928
Dirty Model	3.666^{**}	$3,942^{***}$	0.930
LRA	3.764^{**}	$3,984^{***}$	0.933
MRD	4.472^{***}	$8,041^{***}$	0.885
HI-VAE	4.892^{***}	$10,893^{***}$	0.950
SSHIBA multiple output	$\mathbf{3.406}$	$\mathbf{2,814}$	$\mathbf{0.956}$

Table 5: Results of the simultaneous prediction of three output variables, A, V and D. We used two different scores for this experiment, namely, MAE for A and V and multiclass AUC for D. To display the significance of the permutation test, we use * if $p < 0.05$, ** if $p < 0.01$ and *** if $p < 0.001$.

Table 5 summarises the results obtained with the analysed methods using the respective scoring to measure the performance of the models. These indicate that the proposed approach outperformed the baseline methods in the prediction of the three analysed tasks. The improvement is clearer for A and V, where the proposal outperforms the best baseline by 0.217 and 1,167, respectively, with a high significance in the permutation test. In addition, this performance improvement was achieved while automatically imputing all the missing values in the data, having not only a prediction of the three output variables at month 36 but also for every month where there was no measure. Besides, this performance improvement is accomplished with an effective dimensional reduction, since SSHIBA, calculating a latent representation of the data, is able to outperform the MTL baselines that use all the original features. Looking at the models that find latent representations (MRD and HI-VAE), none of them are capable of capturing in their latent space the temporal nature of the data, while SSHIBA provides outstanding results.

3.4. Analysis of the interpretability

Factor analysis algorithms provide interpretability to the results that may help to identify relations between variables and other information related to the data. Specifically, the sparsity over the latent factors leads to having both common (similar to Canonical Correlation Analysis) and independent (similar to Principal Component Analysis) latent factors between views which, in turn, describes the relationship between variables. In this section, we analyse some of this information learnt by the complete framework.

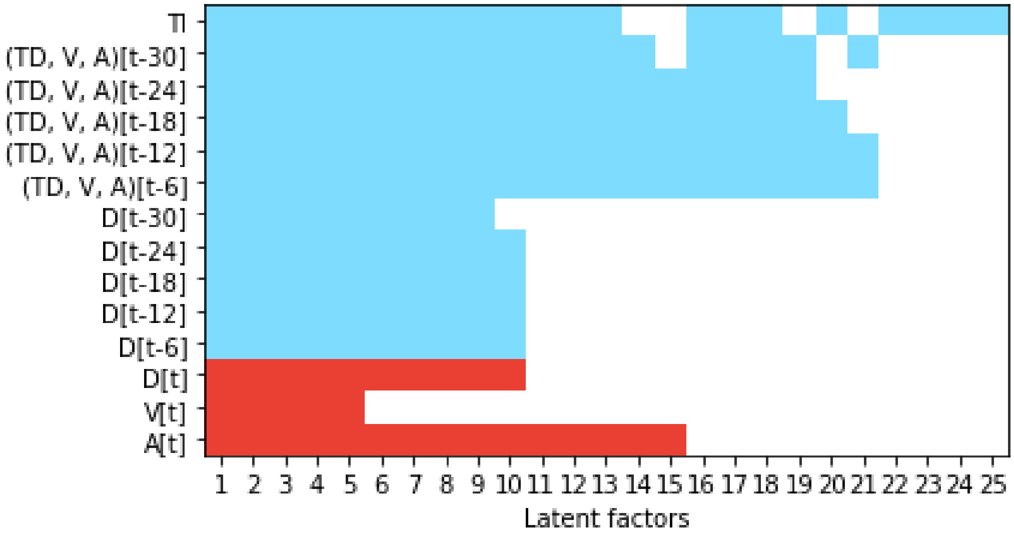


Figure 4: Learnt latent factors with multiple output prediction. Coloured cells imply the latent variable is used to describe the view (blue for input data and red for output), white ones imply the latent variable is not used for this view.

The ARD prior over matrix $\mathbf{W}^{(\mathrm{m})}$ induces zeros in the latent factors, leading to the elimination of some of these irrelevant factors for certain views. In particular, this pruning makes some factors to be common to certain views and not to other, therefore learning the correlation between views. Figure 4 shows the latent factors learnt by the model, where the 14 views have been concatenated for the illustration and the factors have been reordered to show together those with the same active factors for the views. This image demonstrates that the model is learning 5 factors common to all views, where the information of all type of data is combined, another 5 combine all views but do not use the output ventricle information and in one case the diagnosis at the first month. Then, there are 5 views which only use the output information of the ADAS13 score and do not use the information of the diagnosis, two of which do not have the TI variables. Finally, there are 10 latent factors which do not have any output related views and combine the information of the TI and TD variables.

Looking at the needed latent factors for each prediction task, we can see that the prediction of V is the simplest and can be done using just 5 latent factors. Equivalently, D requires 5 more latent factors for the prediction and A is the most complicated task and requires another 5 latent factors, which combine information of TI, TD, V and A.

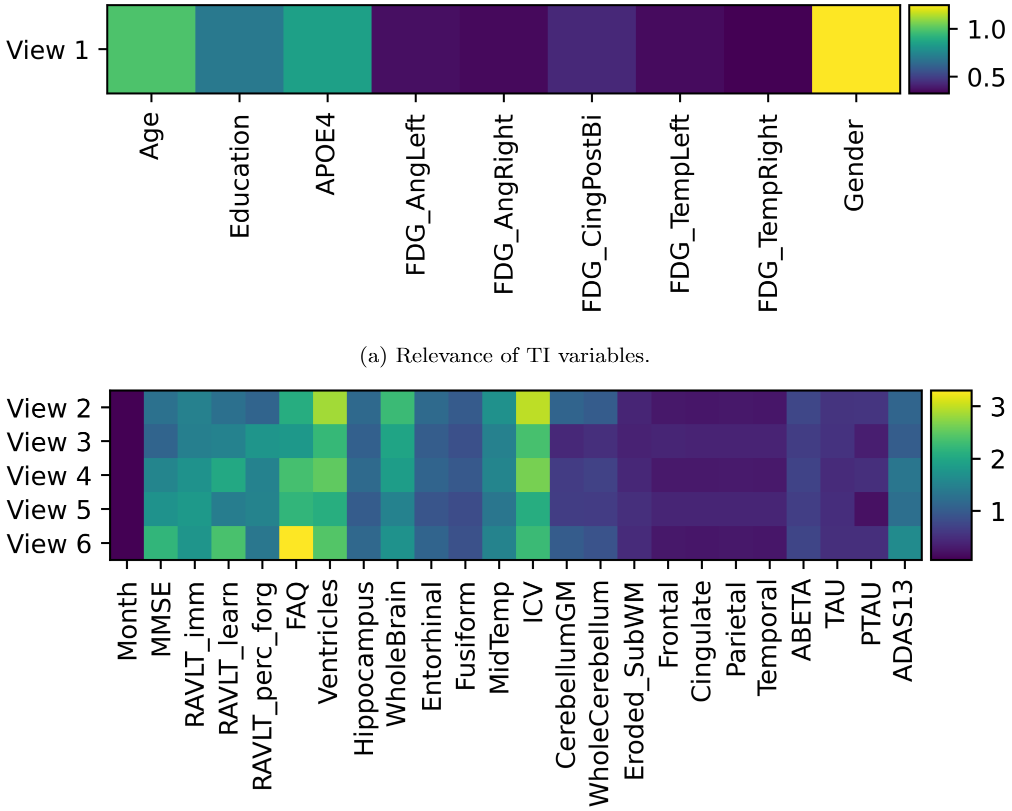
Another functionality of the SSHIBA model is its ability to learn a relevance measure for each input variable of a certain view. The learnt relevances of the model are presented in Figure 5, where we show the relevance learnt for the TI variables (Figure 5a) and the TD variables for multiple time-stamps before the prediction (Figure 5b). These results obtain a higher relevance of the neurophysclogical and behavioural tests as well as the MRI data with respect to the rest of the variables. Furthermore, the difference in scales between both figures demonstrate that the relevance level of the TD variables make the relevance of the TI variables negligible.

However these results are highly correlated to the missing values on each data variable. Looking at the percentage of missing values in Figure 1, we can see the high number of missing values specially in the AVF45 and CSF variables, which leads to a lower feature relevance learnt by the model.

Figure 6 shows some exemplary prediction results obtained with the multioutput SSHIBA framework. These images present the result of the prediction of month 36, while also showing the labels known from previous months (the ones which have a label above the bar) and the imputation of those months for which there was no subject diagnosis (no name above the bar). Looking at Figures 6a and 6c we see that there was a change in diagnosis between the baseline and month 36 which is also learned by the model. Specifically, we can see the label value adapts to the change of clinical diagnosis between timestamps.

4. Discussion

We analysed the viability of the application of SSHIBA for the characterisation of AD based on longitudinal data. The method uses Bayesian variational inference and allows learning the approximate posterior distribution of the model parameters to describe the observations. In this way, the algorithm is able to impute missing values in the observations, functioning in a semi-supervised manner, and combine multiple data information in different views. In particular, it assumes that missing values in the observations correspond to random variables and then computes an approximation to



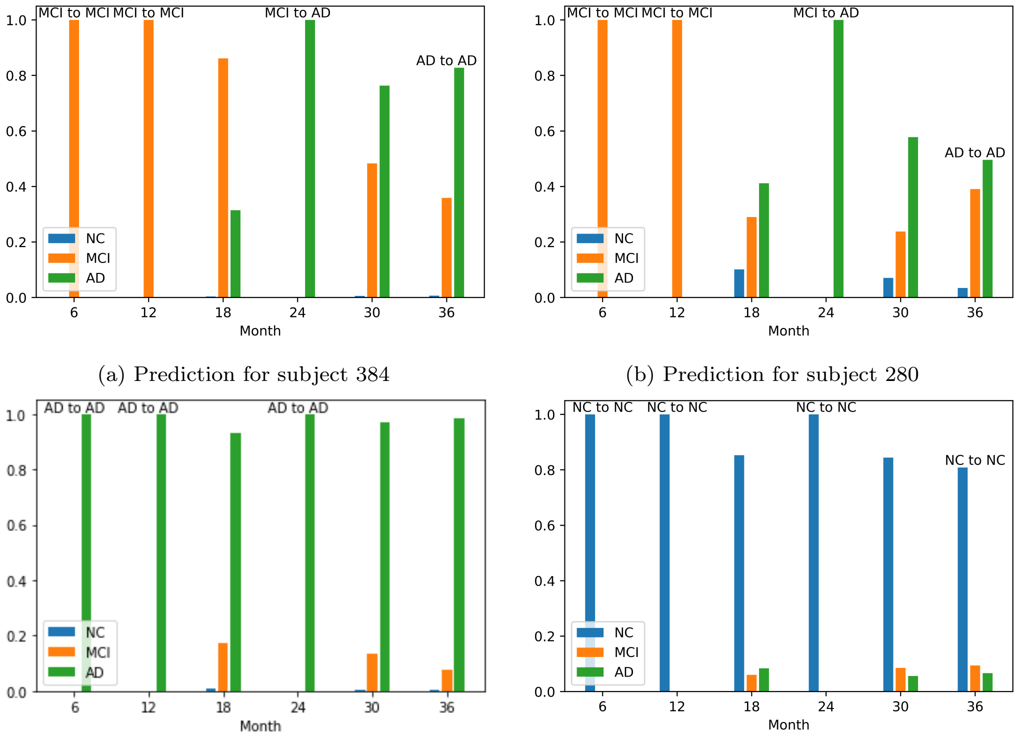
(a) Relevance of TI variables.

(b) Month evolution of the relevance of TD variables.

Figure 5: Analysis of the relevances learnt by the model for each feature. Figure 5a represents all the time-independent variables from the model and Figure 5b the relevance of the time-dependent data for each view.

their true distribution. Once the distribution has been learned, we can either sample or use a statistic, e.g., the mean, to impute unknown values. The multiview formulation allow us to combine data from individual timepoint in a single framework by using a specific projection matrix (weight matrix) for each timepoint. Therefore, the shared latent space combines the distinct data views and learns the relationships between individual timepoints. At the same time, we can use the multiview formulation to model and predict the output variables simultaneously.

The proposed usage of SSHIBA for the characterisation of AD presents a framework that can tackle various usual problems in neuroimaging problems.



(a) Prediction for subject 384

(b) Prediction for subject 280

(c) Prediction for subject 1281

(d) Prediction for subject 649

Figure 6: Exemplary diagnosis prediction for some relevant subjects. The changes of diagnosis of the subject are shown above the bars for each month if available. If the change of diagnosis is not available it shows the predictive probability of each label learnt by the model.

Although there are baselines capable of imputing missing values, modelling longitudinal data, doing multi-task learning or learning feature relevance, SSHIBA poses a breakthrough by being able to simultaneously combining these tasks indistinctly within a single model. This, in turn, allows the model to adequately adapt to the problem needs and combine the information in a reduced latent space.

We used SSHIBA and various baselines to model the TADPOLE database, generated with ADNI data. The results demonstrate how SSHIBA performs in imputing missing values compared to the baselines. The results indicate that SSHIBA achieves more consistent imputation, and an improvement in the prediction accuracy when all available information is combined. Other Bayesian imputation methods like HI-VAE have a hindered performance in this database from the complications of working with temporal data with a high missing value ratio. Furthermore, the comparison with state-of-the-art MTL models show a meaningful outperformance of the proposed framework in the prediction of the three tasks. Specifically, the latent space of the proposed model can efficiently capture temporal relations that considerably improves the performance of both Bayesian latent models and MTL models.

Finally, the Bayesian nature of the formulation provides further information predicting the diagnosis (NC, MCI and AD). Examining the predictions learnt by the model we see that, the algorithm captures the temporal relationships between the variables even though there is no explicitly specified temporal relation in the proposed model. This implies that the model can adapt and assign different weights to the different views (timepoints) so that their combination in the latent space describes their time dependence.

5. Conclusion

In this paper, we have introduced a framework based on a reformulation of the recently proposed SSHIBA model to work in multi-tasks scenarios while dealing with missing values on longitudinal data. Using its multi-view capability, the model combines different time-points with various outputs in a common latent space. We applied the proposed model to jointly predict diagnosis, ventricle volume, and ADAS score in dementia using the ADNI database. The results have proved that the proposed framework is adequate for high missing samples rate scenarios while greatly improving the predictive performance of the three analysed tasks with respect to the baselines. Furthermore, the results have shown that SSHIBA is able to learn the inherent variable time evolution taking advantage of the latent representation of the model.

References

[1] D. E. Barnes and K. Yaffe, “The projected effect of risk factor reduction on alzheimer’s disease prevalence,” The Lancet Neurology, vol. 10, no. 9, pp. 819–828, 2011.

[2] I. Lisko, J. Kulmala, M. Annetorp, T. Ngandu, F. Mangialasche, and M. Kivipelto, “How can dementia and disability be prevented in older adults: where are we today and where are we going?,” Journal of internal medicine, vol. 289, no. 6, pp. 807–830, 2021.

[3] J. K. Kueper, M. Speechley, and M. Montero-Odasso, "The alzheimer's disease assessment scale–cognitive subscale (adas-cog): modifications and responsiveness in pre-dementia populations. a narrative review," Journal of Alzheimer's Disease, vol. 63, no. 2, pp. 423–444, 2018.

[4] J. Sevigny, P. Chiao, T. Bussière, P. H. Weinreb, L. Williams, M. Maier, R. Dunstan, S. Salloway, T. Chen, Y. Ling, et al., "The antibody aducanumab reduces aβ plaques in alzheimer's disease," Nature, vol. 537, no. 7618, pp. 50–56, 2016.

[5] R. Duara, D. Loewenstein, E. Potter, J. Appel, M. Greig, R. Urs, Q. Shen, A. Raj, B. Small, W. Barker, et al., "Medial temporal lobe atrophy on mri scans and the diagnosis of alzheimer disease," Neurology, vol. 71, no. 24, pp. 1986–1992, 2008.

[6] I. Koval, J.-B. Schiratti, A. Routier, M. Bacci, O. Colliot, S. Allassonnière, and S. Durrleman, "Spatiotemporal propagation of the cortical atrophy: Population and individual patterns," Frontiers in neurology, vol. 9, p. 235, 2018.

[7] V. Venkatraghavan, E. J. Vinke, E. E. Bron, W. J. Niessen, M. A. Ikram, S. Klein, M. W. Vernooij, A. D. N. Initiative, et al., "Progression along data-driven disease timelines is predictive of alzheimer's disease in a population-based cohort," NeuroImage, p. 118233, 2021.

[8] M. C. Donohue, H. Jacqmin-Gadda, M. Le Goff, R. G. Thomas, R. Raman, A. C. Gamst, L. A. Beckett, C. R. Jack Jr, M. W. Weiner, J.-F. Dartigues, et al., "Estimating long-term multivariate progression from short-term data," Alzheimer's & Dementia, vol. 10, pp. S400–S410, 2014.

[9] H. M. Fonteijn, M. Modat, M. J. Clarkson, J. Barnes, M. Lehmann, N. Z. Hobbs, R. I. Scahill, S. J. Tabrizi, S. Ourselin, N. C. Fox, et al., "An event-based model for disease progression and its application in familial alzheimer's disease and huntington's disease," NeuroImage, vol. 60, no. 3, pp. 1880–1889, 2012.

[10] S. E. Hardy, H. Allore, and S. A. Studenski, "Missing data: a special challenge in aging research," Journal of the American Geriatrics Society, vol. 57, no. 4, pp. 722–729, 2009.

[11] H. H. Atkinson, C. Rosano, E. M. Simonsick, J. D. Williamson, C. Davis, W. T. Ambrosius, S. R. Rapp, M. Cesari, A. B. Newman, T. B. Harris, et al., "Cognitive function, gait speed decline, and comorbidities: the health, aging and body composition study," The Journals of Gerontology Series A: Biological Sciences and Medical Sciences, vol. 62, no. 8, pp. 844–850, 2007.

[12] G. Martí-Juan, G. Sanroma-Guell, and G. Piella, "A survey on machine and statistical learning for longitudinal analysis of neuroimaging data in alzheimer's disease," Computer methods and programs in biomedicine, vol. 189, p. 105348, 2020.

[13] J. Zhou, J. Liu, V. A. Narayan, J. Ye, A. D. N. Initiative, et al., "Modeling disease progression via multi-task learning," NeuroImage, vol. 78, pp. 233–248, 2013.

[14] M. Huang, W. Yang, Q. Feng, and W. Chen, "Longitudinal measurement and hierarchical classification framework for the prediction of alzheimer's disease," Scientific reports, vol. 7, no. 1, pp. 1–13, 2017.

[15] S. Adhikari, F. Lecci, J. T. Becker, B. W. Junker, L. H. Kuller, O. L. Lopez, and R. J. Tibshirani, "High-dimensional longitudinal classification with the multinomial fused lasso," Statistics in medicine, vol. 38, no. 12, pp. 2184–2205, 2019.

[16] N. McCombe, S. Liu, X. Ding, G. Prasad, M. Bucholc, D. P. Finn, S. Todd, P. L. McClean, and K. Wong-Lin, "Practical strategies for extreme missing data imputation in dementia diagnosis," medRxiv, pp. 2020–07, 2021.

[17] P. Cao, X. Liu, H. Liu, J. Yang, D. Zhao, M. Huang, and O. Zaiane, "Generalized fused group lasso regularized multi-task feature learning for predicting cognitive outcomes in alzheimers disease," Computer methods and programs in biomedicine, vol. 162, pp. 19–45, 2018.

[18] V. Imani, M. Prakash, M. Zare, and J. Tohka, "Comparison of single and multitask learning for predicting cognitive decline based on mri data," IEEE Access, vol. 9, pp. 154275–154291, 2021.

[19] L. Jin, W. Du, B. Ma, D. Zeng, Y. Han, and S. Li, "Feature level-based group lasso method for amnestic mild cognitive impairment diagnosis,"

Computer Methods and Programs in Biomedicine, vol. 208, p. 106286, 2021.

[20] P. Cao, X. Liu, J. Yang, D. Zhao, M. Huang, and O. Zaiane, "ℓ_2, 1-ℓ_1 regularized nonlinear multi-task representation learning based cognitive performance prediction of alzheimer's disease," Pattern Recognition, vol. 79, pp. 195–215, 2018.

[21] P. Yang, F. Zhou, D. Ni, Y. Xu, S. Chen, T. Wang, and B. Lei, "Fused sparse network learning for longitudinal analysis of mild cognitive impairment," IEEE transactions on cybernetics, 2019.

[22] S. Tabarestani, M. Aghili, M. Eslami, M. Cabrerizo, A. Barreto, N. Rishe, R. E. Curiel, D. Loewenstein, R. Duara, and M. Adjouadi, "A distributed multitask multimodal approach for the prediction of alzheimer's disease in a longitudinal study," NeuroImage, vol. 206, p. 116317, 2020.

[23] C. Sevilla-Salcedo, V. Gómez-Verdejo, and P. M. Olmos, "Sparse semi-supervised heterogeneous interbattery bayesian analysis," Pattern Recognition, vol. 120, p. 108141, 2021.

[24] R. V. Marinescu, N. P. Oxtoby, A. L. Young, E. E. Bron, A. W. Toga, M. W. Weiner, F. Barkhof, N. C. Fox, S. Klein, D. C. Alexander, et al., "Tadpole challenge: Prediction of longitudinal evolution in alzheimer's disease," arXiv preprint arXiv:1805.03909, 2018.

[25] M. Prakash, M. Abdelaziz, L. Zhang, B. A. Strange, J. Tohka, A. D. N. Initiative, et al., "Quantitative longitudinal predictions of alzheimer's disease by multi-modal predictive learning," Journal of Alzheimer's Disease, no. Preprint, pp. 1–14, 2020.

[26] R. Duara, W. Barker, R. Lopez-Alberola, D. Loewenstein, L. Grau, D. Gilchrist, S. Sevush, and P. S. George-Hyslop, "Alzheimer's disease: interaction of apolipoprotein e genotype, family history of dementia, gender, education, ethnicity, and age of onset," Neurology, vol. 46, no. 6, pp. 1575–1579, 1996.

[27] S. M. Landau, M. Lu, A. D. Joshi, M. Pontecorvo, M. A. Mintun, J. Q. Trojanowski, L. M. Shaw, W. J. Jagust, and A. D. N. Initiative, "Comparing positron emission tomography imaging and cerebrospinal

fluid measurements of β-amyloid," Annals of neurology, vol. 74, no. 6, pp. 826–836, 2013.

[28] S. M. Landau, D. Harvey, C. M. Madison, R. A. Koeppe, E. M. Reiman, N. L. Foster, M. W. Weiner, W. J. Jagust, A. D. N. Initiative, et al., "Associations between cognitive, functional, and fdg-pet measures of decline in ad and mci," Neurobiology of aging, vol. 32, no. 7, pp. 1207–1218, 2011.

[29] M. Ortner, R. Drost, D. Heddderich, O. Goldhardt, F. Müller-Sarnowski, J. Diehl-Schmid, H. Förstl, I. Yakushev, and T. Grimmer, "Amyloid pet, fdg-pet or mri?-the power of different imaging biomarkers to detect progression of early alzheimer's disease," BMC neurology, vol. 19, no. 1, pp. 1–6, 2019.

[30] M. Samuraki, I. Matsunari, W.-P. Chen, K. Yajima, D. Yanase, A. Fujikawa, N. Takeda, S. Nishimura, H. Matsuda, and M. Yamada, "Partial volume effect-corrected fdg pet and grey matter volume loss in patients with mild alzheimer's disease," European journal of nuclear medicine and molecular imaging, vol. 34, no. 10, pp. 1658–1669, 2007.

[31] P. Battista, C. Salvatore, and I. Castiglioni, "Optimizing neuropsychological assessments for cognitive, behavioral, and functional impairment classification: a machine learning study," Behavioural neurology, vol. 2017, 2017.

[32] M. F. Folstein, S. E. Folstein, and P. R. McHugh, ""mini-mental state": a practical method for grading the cognitive state of patients for the clinician," Journal of psychiatric research, vol. 12, no. 3, pp. 189–198, 1975.

[33] J. Bean, "Rey auditory verbal learning test, rey avlt," Encyclopedia of clinical neuropsychology, pp. 2174–2175, 2011.

[34] R. I. Pfeffer, T. T. Kurosaki, C. Harrah Jr, J. M. Chance, and S. Filos, "Measurement of functional activities in older adults in the community," Journal of gerontology, vol. 37, no. 3, pp. 323–329, 1982.

[35] K. A. Johnson, R. A. Sperling, C. M. Gidicsin, J. S. Carmasin, J. E. Maye, R. E. Coleman, E. M. Reiman, M. N. Sabbagh, C. H. Sadowsky, A. S.

Fleisher, et al., "Florbetapir (f18-av-45) pet to assess amyloid burden in alzheimer's disease dementia, mild cognitive impairment, and normal aging," Alzheimer's & Dementia, vol. 9, no. 5, pp. S72–S83, 2013.

[36] S. M. Landau, M. A. Mintun, A. D. Joshi, R. A. Koeppe, R. C. Petersen, P. S. Aisen, M. W. Weiner, W. J. Jagust, and A. D. N. Initiative, "Amyloid deposition, hypometabolism, and longitudinal cognitive decline," Annals of neurology, vol. 72, no. 4, pp. 578–586, 2012.

[37] S. Shokouhi, J. W. Mckay, S. L. Baker, H. Kang, A. B. Brill, H. E. Gwirtsman, W. R. Riddle, D. O. Claassen, and B. P. Rogers, "Reference tissue normalization in longitudinal 18 f-florbetapir positron emission tomography of late mild cognitive impairment," Alzheimer's research & therapy, vol. 8, no. 1, pp. 1–12, 2016.

[38] L. M. Shaw, H. Vanderstichele, M. Knapik-Czajka, C. M. Clark, P. S. Aisen, R. C. Petersen, K. Blennow, H. Soares, A. Simon, P. Lewczuk, et al., "Cerebrospinal fluid biomarker signature in alzheimer's disease neuroimaging initiative subjects," Annals of neurology, vol. 65, no. 4, pp. 403–413, 2009.

[39] M. Gómez-Sancho, J. Tohka, V. Gómez-Verdejo, A. D. N. Initiative, et al., "Comparison of feature representations in mri-based mci-to-ad conversion prediction," Magnetic resonance imaging, vol. 50, pp. 84–95, 2018.

[40] R. C. Mohs, D. Knopman, R. C. Petersen, S. H. Ferris, C. Ernesto, M. Grundman, M. Sano, L. Bieliauskas, D. Geldmacher, C. Clark, et al., "Development of cognitive instruments for use in clinical trials of antidementia drugs: additions to the alzheimer's disease assessment scale that broaden its scope.," Alzheimer disease and associated disorders, 1997.

[41] N. Raghavan, M. N. Samtani, M. Farnum, E. Yang, G. Novak, M. Grundman, V. Narayan, A. DiBernardo, A. D. N. Initiative, et al., "The adas-cog revisited: novel composite scales based on adas-cog to improve efficiency in mci and early ad trials," Alzheimer's & Dementia, vol. 9, no. 1, pp. S21–S31, 2013.

[42] R. V. Marinescu, N. P. Oxtoby, A. L. Young, E. E. Bron, A. W. Toga, M. W. Weiner, F. Barkhof, N. C. Fox, A. Eshaghi, T. Toni, et al., "The

alzheimer's disease prediction of longitudinal evolution (tadpole) challenge: Results after 1 year follow-up," arXiv preprint arXiv:2002.03419, 2020.

[43] R. M. Neal, Bayesian learning for neural networks, vol. 118. Springer Science & Business Media, 2012.

[44] D. M. Blei, A. Kucukelbir, and J. D. McAuliffe, "Variational inference: A review for statisticians," Journal of the American Statistical Association, vol. 112, no. 518, pp. 859–877, 2017.

[45] R. Tibshirani, "Regression shrinkage and selection via the lasso," Journal of the Royal Statistical Society: Series B (Methodological), vol. 58, no. 1, pp. 267–288, 1996.

[46] A. Evgeniou and M. Pontil, "Multi-task feature learning," Advances in neural information processing systems, vol. 19, p. 41, 2007.

[47] A. Jalali, S. Sanghavi, C. Ruan, and P. Ravikumar, "A dirty model for multi-task learning," Advances in neural information processing systems, vol. 23, pp. 964–972, 2010.

[48] J. Chen, J. Zhou, and J. Ye, "Integrating low-rank and group-sparse structures for robust multi-task learning," in Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 42–50, 2011.

[49] J. Zhou, J. Chen, and J. Ye, "Malsar: Multi-task learning via structural regularization," Arizona State University, vol. 21, 2011.

[50] A. Damianou, N. D. Lawrence, and C. H. Ek, "Multi-view learning as a nonparametric nonlinear inter-battery factor analysis," Journal of Machine Learning Research, vol. 22, no. 86, pp. 1–51, 2021.

[51] A. Nazabal, P. M. Olmos, Z. Ghahramani, and I. Valera, "Handling incomplete heterogeneous data using vaes," Pattern Recognition, vol. 107, p. 107501, 2020.

[52] J. Zhang, Y. Wang, Y. Sun, and G. Li, "Strength of ensemble learning in multiclass classification of rockburst intensity," International Journal for Numerical and Analytical Methods in Geomechanics, vol. 44, no. 13, pp. 1833–1853, 2020.

Acknowledgments

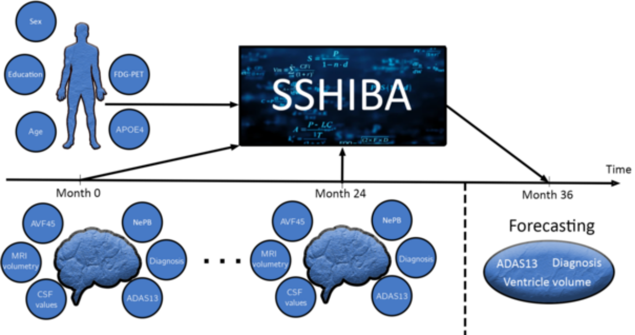
This work made use of the TADPOLE data sets https://tadpole.grand-challenge.org constructed by the EuroPOND consortium http://europond.eu funded by the European Union's Horizon 2020 research and innovation program under grant agreement No 666992. This work was supported by Spanish MINECO (Agencia Estatal de Investigación) [RTI2018-099655-B-100, PID2021-123182OB-I00, TED2021-13182313-100 to P.O.]; and Comunidad de Madrid [IND2018/TIC-9649, Y2018/TCS-4705 to P.O.]; and the Academy of Finland project 316258, "Predictive Brain Image Analysis", to J.T. The work has been performed under the Project HPC-EUROPA3 (INFRAIA-2016-1-730897), with the support of the EC Research Innovation Action under the H2020 Programme to C.S.; in particular, the author gratefully acknowledges the support of A.I. Virtanen Institute for Molecular Sciences and the computer resources and technical support provided by CSC – IT Center for Science. 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 (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.

We thank Mithilesh Prakash for his help in data processing as well as commenting on early drafts of this manuscript.

Conflict of interest statement

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.


Sex
Education
Age
FDG-PET
APOE4
SSHIBA
Time
Month 0
Month 24
Month 36
AVF45
NePB
MRI volumetry
Diagnosis
CSF values
ADAS13
AVF45
NePB
MRI volumetry
Diagnosis
CSF values
ADAS13
Forecasting
ADAS13
Diagnosis
Ventricle volume