Combining Multi-Task Learning and Multi-Channel Variational Auto-Encoders to Exploit Datasets with Missing Observations -Application to Multi-Modal Neuroimaging Studies in Dementia

Luigi Antelmi, Nicholas Ayache, Philippe Robert, Federica Ribaldi, Valentina Garibotto, Giovanni B Frisoni, Marco Lorenzi

Combining Multi-Task Learning and Multi-Channel Variational Auto-Encoders to Exploit Datasets with Missing Observations - Application to Multi-Modal Neuroimaging Studies in Dementia

Luigi Antelmi[1,*], Nicholas Ayache[1], Philippe Robert[2,3], Federica Ribaldi[4,5,6,7], Valentina Garibotto[8,9], Giovanni B. Frisoni[6,7], Marco Lorenzi[1], for the Alzheimer's Disease Neuroimaging Initiative[**]

Abstract

The joint modeling of neuroimaging data across multiple datasets requires to consistently analyze high-dimensional and heterogeneous information in presence of often non-overlapping sets of views across data samples (e.g. imaging data, clinical scores, biological measurements). This analysis is associated with the problem of missing information across datasets, which can happen in two forms: missing at random (MAR), when the absence of a view is unpredictable and does not depend on the dataset (e.g. due to data corruption); missing not at random (MNAR), when a specific view is absent by design for a specific dataset. In order to take advantage of the increased variability and sample size when pooling together observations from many cohorts and at the same time cope with the ubiquitous problem of missing information, we propose here a multi-task generative latent-variable model where the common variability across datasets stems from the estimation of a shared latent representation across views. Our formulation allows to retrieve a consistent latent representation common to all views and datasets, even in the presence of missing information. Simulations on synthetic data show that our method is able to identify a common latent representation of multi-view datasets, even when the compatibility across datasets is minimal. When jointly analyzing multi-modal neuroimaging and clinical data from real independent dementia studies, our model is able to mitigate the absence of modalities without having to discard any available information. Moreover, the common latent representation inferred with our model can be used to define robust classifiers gathering the combined information across different datasets. To conclude, both on synthetic and real data experiments, our model compared favorably to state of the art benchmark methods, providing a more powerful exploitation of multi-modal observations with missing views.

1. Introduction

Because of the inherent complexity of biomedical data and diseases, researchers are required to integrate data across different studies to increase the sample size and obtain better models (Le Sueur et al., 2020). In the development of integrative models, researchers have to face with multiple concurrent challenges, such as the ones related to datasets interoperability (Tognin et al., 2020), data heterogeneity (Buch and Liston, 2020), and data missingness (Golriz Khatami et al., 2020). Emblematic is the case of integrative modeling when datasets come from multi-centric studies in cognitive and neurological disorders, such as in Alzheimer's Disease (AD). Here the datasets interoperability is hampered by the existence of different protocols between studies. Because of this, methods whose modeling task are specifically designed on one dataset cannot be directly applied to another one. Furthermore, at the level of each single dataset, researchers face the challenge of modeling heterogeneous data, such as multiple imaging modalities, clinical scores and biological measurements. Because each one of these sources of information represents an important and independent *view* on the disease or phenomena under investigation, efforts to model multi-view data are ubiquitous in the recent literature (Vieira et al., 2020), where the objective ranges from predicting clinical outcomes (Chen et al., 2019) to synthesizing new modalities (Zhou et al., 2020; Wei et al., 2019). Another common problem to the joint modeling of multiple datasets is represented by missing data. At the level of the single datasets, views can be missing at random (MAR) for some subjects. Typ-

[*]Corresponding author: luigi.antelmi@inria.fr

[**]Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (http://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://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf

[1]University of Côte d'Azur, Inria, Epione Project-Team, France.

[2]University of Côte d'Azur, CoBTeK, France.

[3]Centre Mémoire, CHU of Nice, France

[4]Laboratory of Alzheimer's Neuroimaging and Epidemiology (LANE), Saint John of God Clinical Research Centre, Brescia, Italy.

[5]Department of Molecular and Translational Medicine, University of Brescia, Brescia, Italy.

[6]Laboratory of Neuroimaging of Aging (LANVIE), University of Geneva, Geneva, Switzerland.

[7]Memory Clinic, Geneva University Hospitals, Geneva, Switzerland.

[8]Laboratory of Neuroimaging and Innovative Molecular Tracers (NIMTlab), Geneva University Neurocenter and Faculty of Medicine, University of Geneva, Geneva, Switzerland.

[9]Division of Nuclear Medicine and Molecular Imaging, Diagnostic Department, Geneva University Hospitals, Geneva, Switzerland.

ically, as fitting multi-view models requires to establish connections between views, observations with at least one missing view are discarded, yielding to potentially severe loss of available information. To mitigate this problem, imputation methods are usually applied to infer missing views by modeling the relationship across views from complete observations. The loss of information is exacerbated when considering multiple datasets altogether. Indeed, according to the cohort study design, there may be views which are specifically absent, hence missing not at random (MNAR). This potential mismatch across datasets hampers their interoperability, and prevents the gathering of all the available observations into a single, robust and generalizable joint model accounting for the global data variability. This challenge is typically addressed in machine learning by the field of Multi Task Learning (MTL). To address this issue, MTL aims at improving the model interoperative capabilities by exploiting the information extracted from multiple datasets. In MTL each task is usually associated to the modeling of a specific dataset and its views only, and the main idea is to share across datasets the parameters learned through each modeling task (Caruana, 1998; Dorado-Moreno et al., 2020). As an example of MTL, in model-agnostic meta-learning (MAML) (Finn et al., 2017) the training of a model on a variety of learning tasks enforces the generalization on new datasets after few fine tuning iterations. In the context of data assimilation, MTL is usually achieved with specific output layers for every task, and by including a shared representation for all of them (Dorado-Moreno et al., 2020). This modeling rationale is at the basis of recent MTL based approaches to heterogeneous data assimilation (Wu and Goodman, 2018; Antelmi et al., 2019; Shi et al., 2019). In particular, approaches such as the The Multi-Channel Variational Autoencoder (MCVAE) (Antelmi et al., 2019) rely on the identification of a common latent representation for different views belonging to a single dataset. Training these models is possible with some limitations: 1) after having discarded observations with missing views; 2) when all the training observations are compatible in terms of available views, and hence are usually limited to model one dataset at a time.

To overcome these limitations, here we investigate an extension of MCVAE to simultaneously learn from multiple datasets, even in the presence of non compatible views between datasets, and missing views within datasets. Our extension (Fig. 2) is built upon the following actions: 1) To define tasks across datasets based on the identification of data subsets presenting compatible views, 2) to stack multiple instances of the MCVAE, where each instance models a specific task, 3) share the models parameters of common views between modeling tasks. Thanks to these actions, our method allows to learn a joint model for all the subjects without discarding any information. The common views between tasks act as a bridge and enable the information to flow through all the other views. In the training phase, tasks lacking a particular view will simply not contribute to the learning of those view-specific parameters. All the tasks will nevertheless benefit from the parameters they didn't contribute to learn, for the prediction of their missing views. The proposed variational formulation for computing approximate posterior distributions of the latent variables allows fast and scalable training. Being dataset agnostic, our method allows to integrate all the available data into a joint model, gathering all the available information from multiple datasets at the same time.

The rest of this paper is structured as follow. In § 2 we set the theoretical framework for our model. In § 3 experiments on synthetic data show that the prediction error of missing views is competitive with respect to the one obtained with state of the art imputation methods. In § 4 experiments on real data from independent multi-modal neuroimaging datasets show that our model generalizes better than dataset-specific models on new unseen data, in both the tested cases of data reconstruction and diagnosis classification. Lastly we discuss our results and conclude our work with summary remarks.

2. Method

In this section we recall the theoretical framework of the *Multi-Channel Variational Autoencoder* (MCVAE) developed in our previous work (Antelmi et al., 2019), which we now extend to tackle the problem of missing data integration. In § 2.1 and § 2.2 we introduce the approach and derive the model in presence of missing data. In § 2.3 we briefly recall the main parametric functions adopted later in our experiments with missing data. In § 2.4 we finally propose the new optimization scheme allowing to account for observations with partially missing views. Code developed in Pytorch (Paszke et al., 2019) is publicly available at `https://gitlab.inria.fr/epione_ML/mcvae`.

2.1. Generative Model

Let $\mathcal{D} = \{D_d\}_{d=1}^{D}$ be a collection of D independent datasets, where each dataset $D_d = \{\mathbf{x}_{d,n}\}_{n=1}^{N_d}$ is composed by N_d independent data-points (*e.g.,* subjects in the case of medical imaging datasets). Every dataset D_d is associated with a total number of V_d available views (*e.g.,* sets of clinical scores and imaging derived phenotypes extracted from multiple imaging modalities), and we assume that each data-point $\mathbf{x}_{d,n} = \{\mathbf{x}_{d,n,v}\}_{v=1}^{V_{d,n}}$ is composed by $V_{d,n}$ views, where $V_{d,n} \leq V_d$. With the latest inequality we account for data-points with an arbitrary number of missing views.

For each view $\mathbf{x}_{d,n,v}$ we rely on the following generative latent variable model:

$$\begin{aligned} &\mathbf{z}_{d,n} \sim p(\mathbf{z}), \\ &\mathbf{x}_{d,n,v} \sim p(\mathbf{x}_{d,n,v}|\mathbf{z}_{d,n}, \boldsymbol{\theta}_v), \qquad \text{for } v \text{ in } 1 \dots V_{d,n} \leq V_d, \end{aligned} \tag{1}$$

where $p(\mathbf{z})$ is a prior distribution for the latent variable $\mathbf{z}_{d,n}$ commonly shared by the $V_{d,n}$ views, and where the likelihood functions $p(\mathbf{x}_{d,n,v}|\mathbf{z}_{d,n}, \boldsymbol{\theta}_v)$ belong to a family of distributions parametrized by $\boldsymbol{\theta}_v$, which represents the view-specific generative parameters shared among all datasets.

2.2. Variational Inference

The exact solution to the inference problem is given by the posterior $p\left(\mathbf{z}|\{\mathbf{x}_{d,n,v}, \boldsymbol{\theta}_v\}_{v=1}^{V_{d,n}}\right)$, that is not generally computable analytically. Following Antelmi et al. (2019), we can nevertheless look for its approximation $q(\mathbf{z})$ through *Variational*

Inference (Blei et al., 2017). By introducing the latent variational approximation $q(\mathbf{z})$, we can derive the lower bound on the marginal log-likelihood for a single data-point as follows:

$$\begin{aligned}\ln p(\mathbf{x}_{d,n,v}|\boldsymbol{\theta}_v) &= \ln \int p(\mathbf{x}_{d,n,v}|\mathbf{z},\boldsymbol{\theta}_v)\, p(\mathbf{z})\ d\mathbf{z} \\ &= \ln \int \frac{q(\mathbf{z})}{q(\mathbf{z})} p(\mathbf{x}_{d,n,v}|\mathbf{z},\boldsymbol{\theta}_v)\, p(\mathbf{z})\ d\mathbf{z} \\ &= \ln \mathbb{E}_{q(\mathbf{z})}\left[\frac{p(\mathbf{x}_{d,n,v}|\mathbf{z},\boldsymbol{\theta}_v)\, p(\mathbf{z})}{q(\mathbf{z})}\right] \\ &\geq \mathbb{E}_{q(\mathbf{z})}\left[\ln p(\mathbf{x}_{d,n,v}|\mathbf{z},\boldsymbol{\theta}_v)\right] - \mathcal{D}_{\mathrm{KL}}\left(q(\mathbf{z})\|p(\mathbf{z})\right).\end{aligned} \tag{2}$$

To derive the last line of Eq. (2) we leverage on the *Jensen's inequality* and collect the result into a new expectation term and in the Kullback-Leibler divergence term ($\mathcal{D}_{\mathrm{KL}}$).

We define the distribution function $q(\mathbf{z})$ to depend on a specific dataset d, data-point n, and view w, such that:

$$q(\mathbf{z}) = q_{d,n,w}(\mathbf{z}) = q(\mathbf{z}|\mathbf{x}_{d,n,w},\boldsymbol{\phi}_w), \tag{3}$$

where $\boldsymbol{\phi}_w$ represents the view-specific variational parameters shared among all datasets. To force a link among views, we impose the inequality Eq. (2) to hold for any w in $1 \dots V_{d,n}$. To do so, we average the right hand side of Eq. (2) across the $V_{d,n}$ views and rewrite Eq. (2) as follows:

$$\ln p(\mathbf{x}_{d,n,v}|\boldsymbol{\theta}_v) \geq \mathcal{L}_v^{(\mathbf{x}_{d,n})} = \frac{1}{V_{d,n}} \sum_{w=1}^{V_{d,n}} \mathcal{L}_{v\leftarrow w}^{(\mathbf{x}_{d,n})}, \tag{4}$$

where

$$\mathcal{L}_{v\leftarrow w}^{(\mathbf{x}_{d,n})} = \mathbb{E}_{q_{d,n,w}(\mathbf{z})}\left[\ln p(\mathbf{x}_{d,n,v}|\mathbf{z},\boldsymbol{\theta}_v)\right] - \mathcal{D}_{\mathrm{KL}}\left(q_{d,n,w}(\mathbf{z})\|p(\mathbf{z})\right) \tag{5}$$

is the lower bound associated to the data-point $\mathbf{x}_{d,n}$ when its view v is predicted from its view w. In Fig. 1 we sketch the model structure induced by Eq. (5).

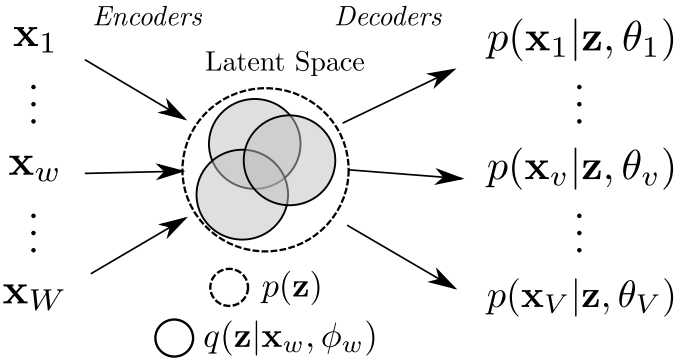


Figure 1: General variational framework for our multi-view model. For every pair of views v and w there is a prediction path $v \leftarrow w$ composed by two learnable functions: the encoding distribution $q(\mathbf{z}|\mathbf{x}_w,\boldsymbol{\phi}_w)$ and the decoding likelihood $p(\mathbf{x}_v|\mathbf{z},\boldsymbol{\theta}_v)$. Parameters $\boldsymbol{\phi}_w$ and $\boldsymbol{\theta}_v$ are optimized through Eq. (9) to maximize the likelihood of our generative model under the encoding distributions, and at the same time minimize the Kullback-Leibler distance between every encoder and the prior $p(\mathbf{z})$. To leave the notation uncluttered, in this representation we dropped the dataset index d and data-point n used in the text.

2.3. Parameterization

With the right choice of the functional form of $q(\mathbf{z}|\mathbf{x}_{d,n,w},\boldsymbol{\phi}_w)$, $p(\mathbf{z})$, and $p(\mathbf{x}_{d,n,v}|\mathbf{z},\boldsymbol{\theta}_v)$, the right hand side of Eq. (4) becomes amenable to computation and optimization, yielding to the maximization of the left hand side, quantity also known as the model evidence. Of course, the choice for the likelihood function $p(\mathbf{x}_{d,n,v}|\mathbf{z},\boldsymbol{\theta}_v)$ depends on the nature of the view $\mathbf{x}_{d,n,v}$. For example it can be parametrized as a multivariate Gaussian in the case of continuous data (*i.e.,*imaging derived phenotypes), as a Bernoulli likelihood for dichotomic data, and as a Categorical likelihood for categorical data.

In general, the prior distribution $p(\mathbf{z})$ is the multivariate Gaussian distribution $\mathcal{N}(\mathbf{0};\mathbf{I})$. The same family of distributions is also commonly used for the variational and likelihood functions, such that respectively:

$$q(\mathbf{z}|\mathbf{x}_{d,n,w},\boldsymbol{\phi}_w) = \mathcal{N}\left(\boldsymbol{\mu} = \mathbf{V}_w^{(\mu)}\mathbf{x}_{d,n,w}\,;\, \boldsymbol{\Sigma} = \mathrm{diag}\left(\mathbf{V}_w^{(\sigma)}\mathbf{x}_{d,n,w}\right)\right), \tag{6}$$

$$p(\mathbf{x}_{d,n,v}|\mathbf{z},\boldsymbol{\theta}_v) = \mathcal{N}\left(\boldsymbol{\mu} = \mathbf{G}_v^{(\mu)}\mathbf{z} \quad ;\, \boldsymbol{\Sigma} = \mathrm{diag}\left(\mathbf{g}_v^{(\sigma)}\right)\right), \tag{7}$$

where the moments $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are obtained from linear transformations of the conditioning variables. Here, $\boldsymbol{\theta}_v = \{\mathbf{G}_v^{(\mu)}, \mathbf{g}_v^{(\sigma)}\}$ and $\boldsymbol{\phi}_w = \{\mathbf{V}_w^{(\mu)}, \mathbf{V}_w^{(\sigma)}\}$ are the parameters to be optimized. A non-linear parameterization can be used as well, for example in the form of deep neural networks.

In Antelmi et al. (2019) we also introduced the following alternative parameterization for the posterior distribution:

$$q_{d,n,w}(\mathbf{z}) = \mathcal{N}\left(\boldsymbol{\mu} = \mathbf{V}_w^{(\mu)}\mathbf{x}_{d,n,w}; \boldsymbol{\Sigma} = \mathrm{diag}\left(\sqrt{\boldsymbol{\alpha}} \odot \boldsymbol{\mu}^2\right)\right), \tag{8}$$

which is known as *dropout posterior* (Kingma et al., 2015). The dropout parameter $\boldsymbol{\alpha}$ has components $\alpha_i = p_i/1-p_i$ linked to the probability p_i of dropping out the i-th latent variable component (Wang and Manning, 2013). It has been shown that the association of this dropout posterior with a log-uniform prior distribution $p(\mathbf{z})$ leads to sparse and interpretable models (Antelmi et al., 2019; Molchanov et al., 2017).

2.4. Optimization

Assuming independent observations, the marginal log-likelihood in the left hand side of Eq. (4) can be summed up over all the datasets, data-points, and views. As a consequence, inference on the model generative parameters $\boldsymbol{\theta} = \{\boldsymbol{\theta}_v\}$ and variational parameters $\boldsymbol{\phi} = \{\boldsymbol{\phi}_w\}$ can be achieved by solving the maximization problem:

$$\begin{aligned}\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\phi}} &= \arg\max_{\boldsymbol{\theta},\boldsymbol{\phi}} \sum_{d,n,v} \mathcal{L}_v^{(\mathbf{x}_{d,n})} \\ &= \arg\max_{\boldsymbol{\theta},\boldsymbol{\phi}} \sum_{d,n,v} \frac{1}{V_{d,n}} \sum_{w=1}^{V_{d,n}} \mathcal{L}_{v\leftarrow w}^{(\mathbf{x}_{d,n})}.\end{aligned} \tag{9}$$

We use Algorithm 1 to solve Eq. (9). The summation in Eq. (9) is done for every dataset d along all the available data-points n and their specific views v. We note that batching data-points with common views can speed up the computation by reducing the number of second level *for* loop iterations in Algorithm 1.

Algorithm 1 Multi-view model optimization.

Require:
Set the dimensionality of $\mathbf{z}$.
Initialize the model parameters $\boldsymbol{\phi}, \boldsymbol{\theta}$.
Set the optimizer learning rate.

while $\boldsymbol{\phi}, \boldsymbol{\theta}$ not converged **do**
  Initialize the total cost:
  $\mathcal{L} \leftarrow 0$
  for every dataset $d \in D$ **do**
    for every datapoint $\mathbf{x}_{d,n}, n \in N_d$ **do**
      for every view $v \in V_{d,n}$ **do**
        Accumulate the cost of predicting v from w:
        $\mathcal{L}_v \leftarrow 0$
        for every view $w \in V_{d,n}$ **do**
          $\mathcal{L}_v \leftarrow \mathcal{L}_v + \mathcal{L}_{v \leftarrow w}^{(\mathbf{x}_{d,n})}$. See Eq. (5).
        end for
        Accumulate the average $\mathcal{L}_v$ in the total cost:
        $\mathcal{L} \leftarrow \mathcal{L} + \frac{1}{V_{d,n}} \mathcal{L}_v$.
      end for
    end for
  end for
  $\boldsymbol{\theta}, \boldsymbol{\phi} = \text{Optim}(\boldsymbol{\phi}, \boldsymbol{\theta}, \nabla_{\boldsymbol{\phi}} \mathcal{L}, \nabla_{\boldsymbol{\theta}} \mathcal{L})$. *Adam* optimizer used to maximize $\mathcal{L}$.
end while

If missing, a particular view v will be simply not accounted for that specific observation, without having to discard all the other views that can still contribute to optimize Eq. (9). The presence of at least one common view among datasets acts as a link across datsets and allows the information to flow through all the datasets to the other views. In Fig. 2 we show the learning scheme of our model in a simple case with four views and one common view between batches.

We observe that this model extends the Multi-Channel VAE (Antelmi et al., 2019), which is a multi-view extension of the VAE (Kingma and Welling, 2014; Rezende et al., 2014). In the former work we proposed a multi-view generative model where they require all the observation in the training set to have all the available views, hence limited to model one dataset at a time (in the case of datasets with different views), after having discarded incomplete observations in that dataset. We address this limitation by allowing missing views in the training set for some observations. The trained model can estimate missing views $\hat{\mathbf{x}}_{d,n,v}$ from the available ones through the formula:

$$\hat{\mathbf{x}}_{d,n,v} = \frac{1}{V_{d,n} - 1} \sum_{w=1, w \neq v}^{V_{d,n}} \mathbb{E}_{q_{d,n,w}(\mathbf{z})} \left[p\left(\mathbf{x}_{d,n,v} | \mathbf{z}, \boldsymbol{\theta}_v\right) \right], \tag{10}$$

where the available views $\mathbf{x}_{d,n,w}$ are encoded into the distributions $q_{d,n,w}$, which are then used to predict the missing view through its decoding distribution $p\left(\mathbf{x}_{d,n,v} | \mathbf{z}, \boldsymbol{\theta}_v\right)$.

3. Synthetic Experiments

In this section we describe our results on extensive synthetic experiments performed with our model and different benchmark methods in two conditions: 1) missing at random views for each dataset, and 2) datasets with systematically missing views (missing not at random).

3.1. Data preparation

To simulate multi dataset observations, we sample the latent variable $\mathbf{z}_{d,n}$ from a multivariate Gaussian with zero-mean and identity covariance matrix, and subsequently we transform each sample with random linear mapping towards the observation space to obtain $\mathbf{x}_{d,n,v}$. We then corrupt the observations with increasing levels of noise and we finally remove views in the context of the *missing at random* (MAR) and *missing not at random* (MNAR) experiments.

In the MAR experiments views were randomly removed according to a parameter $0 \leq f \leq 1$, which controls the fraction of data-points with complete views. In the limit case $f = 1$, each data-point has all the views, representing the ideal case of no missing views, that is the working case of the Multi-Channel Variational Autoencoder (Antelmi et al., 2019). In the case $f = 0$, each data-point has one and only one randomly assigned view, representing the extreme case where no direct relationship between views is observable. Here our multi-view model collapses into a disjoint series of independent Variational Autoencoders (Kingma and Welling, 2014; Rezende et al., 2014). In the general case, each data-point has probability f to have all the views, and probability $1 - f$ to have a randomly assigned view out of the total available views. The general case represents the case where the relationship between views can be established only through a fraction f of the total available data-points.

In the MNAR experiments we removed specific views for each simulated dataset, ensuring at the same time the absence of at least one view for a datasets, and the presence of at least one view in common between pairs of datasets. As an example, in the case with three datasets and three views, the association view-dataset can be expressed through the following association matrix A:

$$A = \begin{pmatrix} 1 & 0 & 1 \\ 1 & 1 & 0 \\ 0 & 1 & 1 \end{pmatrix}, \tag{11}$$

where $A(v, d) = 1$ indicates the presence of view v in dataset d. For experimental purposes we limited our MNAR simulations to cases that can be defined with square association matrices having a dimensionality not greater than 5×5.

3.2. Model Fitting and Evaluation

In both MAR and MNAR experiments we fit the synthetic scenarios with our model, where we choose a linear Gaussian parametrization for variational and likelihood distributions, made explicit respectively in Eq. (6) and Eq. (7). For each simulated scenario we predicted the missing views according to Eq. (10) on testing hold-out datasets.

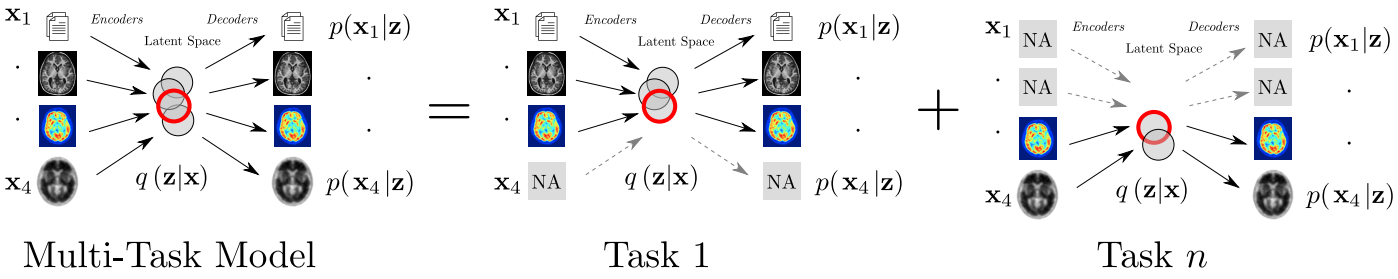


Figure 2: Simple example of a Multi-Task Model learning scheme in the presence of missing not available (NA) views. Arrows represent learnable functions used as network encoders and decoders, transforming respectively input views (*e.g.,* clinical scores, imaging derived phenotypes, ...) from the observation space to the representation space (circles) and from the representation space back to the observation space. The separability of the loss function $\mathcal{L}_v^{(\mathbf{x}_{d,n})}$ in Eq. (4) allows to group together observations into homogeneous learning tasks. For every task, functions associated to missing views (dashed gray arrows) are locally not updated by the learning algorithm. Globally, common latent representations (red circles) across pairs of tasks act as a link allowing the information to flow throughout the views.

Results, cross-validated on 5 folds, are summarized with the *mean squared error* (MSE) metric on testing hold-out datasets for every simulated scenario. We applied the same evaluation procedure for the benchmark methods.

3.3. Benchmark Methods

Among state of the art multivariate linear and non linear imputation methods, we selected the following benchmark approaches: 1) k-Nearest Neighbors (KNN) with $k = \{1, 5\}$; 2) Denoising Autoencoder (DAE) (Vincent et al., 2008); 3) Multivariate Imputation by Chained Equations (MICE) (van Buuren and Groothuis-Oudshoorn, 2000).

For the KNN approach we used the *KNNImputer* method as implemented in the *Scikit-Learn* library (Pedregosa et al., 2011). Here each sample's missing values are imputed using the mean value from k nearest neighbors found in the training set, according to their Euclidean distance.

The Denoising Autoencoder, as developed by Gondara and Wang (2018), is based on an overcomplete deep autoencoder. It maps input data to a higher dimensional space which, in combination with an initial dropout layer inducing corruption, makes the model robust to missing data. We used the same architecture proposed by the authors, that is three hidden layers for encoder and decoder networks, Tanh activation functions, hyperparameter $\Theta = 7$, and dropout $p = 0.5$, as they proved to provide consistently better results.

In MICE, as implemented in van Buuren and Groothuis-Oudshoorn (2011), missing values are modeled as a multivariate linear combination of the available features. This methodology is attractive if the multivariate distribution is a reasonable description of the data, which in our case it is by construction. MICE specifies the multivariate imputation model on a variable-by-variable basis by a set of conditional densities, one for each incomplete variable. Starting from an initial imputation, MICE draws imputations by iterating over the conditional densities.

3.4. Results

In the synthetic tests our model provides the best performances overall, with a mean MSE improvement compared to the best competing method of 17% in MAR cases and 71% in MNAR cases (Fig. 3). We notice that DAE is not always better than KNN ($k = 5$), especially in low SNR cases. We were able to fit the MICE model only on MNAR cases with high SNR where it performed poorly (boxplot not shown), while in all the other cases, including all MAR cases, the model did not converge.

In Fig. 4 we show MAR experiments results stratified by SNR and by the fraction f of data-points with complete views. Here we notice how with already $f = 0.25$ we can significantly reduce the prediction error on testing data-points compared to the case $f = 0$, where no relationship between views can be established. Moreover, reaching the ideal case of $f = 1$, that is when there are no missing views in the dataset, does not improve significantly the prediction performance of our model compared to the case $f = 0.25$.

4. Experiments on Medical Imaging Datasets

In this section we describe our results on jointly modeling real medical imaging datasets, independently acquired in the context of dementia studies.

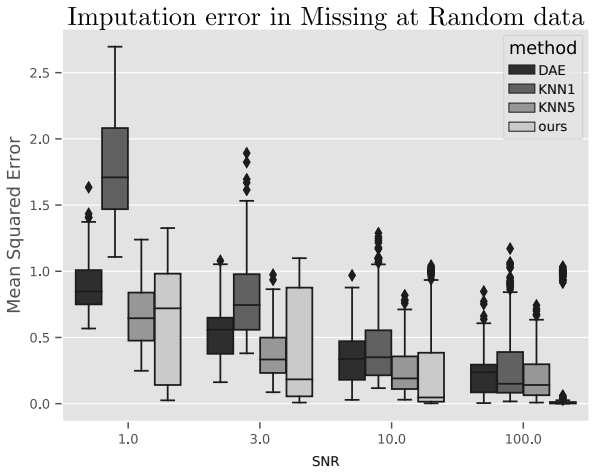
We executed three kinds of experiments: 1) benchmark evaluation of our model against the best competing methods from the previous section; 2) multi-view feature prediction with our model on all the available datasets in multiple training and testing conditions. 3) diagnosis classification from multi-view heterogeneous data in different training and testing conditions.

4.1. Data Sources

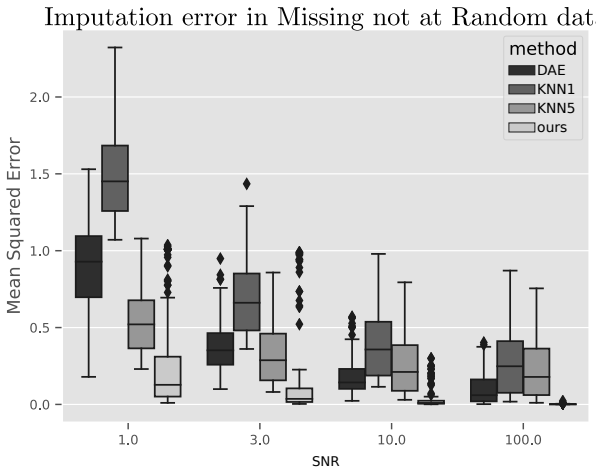
Data used in the preparation of this section were obtained from the following sources.

1. The Alzheimer's Disease Neuroimaging Initiative (ADNI) [10], a database of brain imaging and related clinical data of cognitively normal subjects, and on patients presenting various degrees of cognitive decline.
2. MIRIAD dataset, a database of brain imaging and related clinical data of cognitively normal subjects and patients affected by Alzheimer's disease (Malone et al., 2013).

[10] adni.loni.usc.edu. The ADNI was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD. For up-to-date information, see www.adni-info.org.



(a)



(b)

Figure 3: Mean Squared Error (MSE) of imputation in synthetic held-out datasets (5-folds cross-validation). Compared to the best competing methods among k-Nearest Neighbor ($k = \{1, 5\}$) and Denoise Autoencoder (DAE), our model comes out as the best performer, with a mean MSE improvement of 17% in MAR cases (a) and 71% in MNAR cases (b). Stratification by signal-to-noise ratio (SNR) is shown.

3. OASIS-3 dataset, a database of brain imaging and related clinical data of cognitively normal subjects and subjects at various stages of cognitive decline (LaMontagne et al., 2019).
4. A local cohort collected at the Geneva University Hospitals, with brain imaging and related clinical data of patients with various cognitive disorders.

Subjects enrollment, data collection, and data sharing were approved by the ethic committees associated to each study dataset in accordance with the principles of the Declaration of Helsinki.

The available imaging modalities comes from the following acquisitions:

1. structural Magnetic Resonance Imaging (MRI) to measure anatomical volumes in the brain;

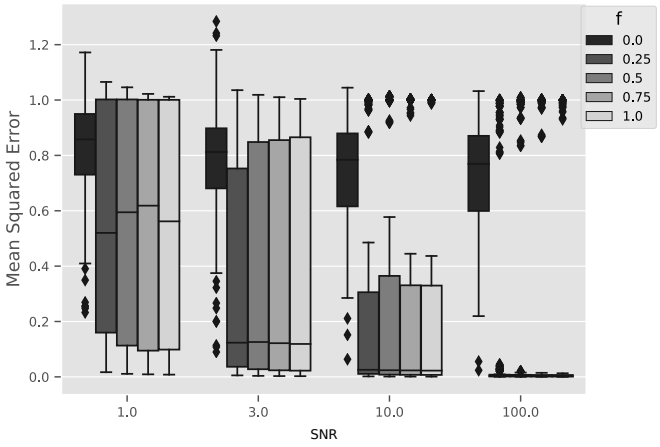


Figure 4: Mean Squared Error of test sets predictions in synthetic held-out datasets in MAR scenarios. Stratification by SNRand by the fraction f of data-points with complete views is shown. A value of $f = 0.25$ is enough to reduce the prediction error on testing data-points at the level of the ideal case ($f = 1$).

2. Positron Emission Tomography (PET) with Fluorodeoxyglucose (FDG) tracer, to measure the glucose uptake, which reflects the functional status of the brain;
3. PET with the AV45 tracer, to measure the amyloid deposits in the brain;
4. PET with the AV1451 tracer, to measure the tau protein aggregates in the brain.

We divided the ADNI dataset into two complementary datasets: 'Adni1', composed by subjects recruited in the initial ADNI1 study (2004-2009), and 'Adni2' composed by those subjects subsequently recruited in ADNI-GO, ADNI2, and ADNI3 (2010-ongoing). Data modalities and acquisition protocols of 'Adni1' present differences from those of 'Adni2'. Specifically, in 'Adni1' and 'Adni2' the MRI imaging was performed respectively on 1.5T and 3T scanners. The two cohorts differs also for the presence of PET imaging data. Therefore we consider these two cohorts as separated datasets.

To summarize, we grouped our data into five distinct datasets which we named as follows: 'Adni1', 'Adni2', 'Miriad', 'Oasis3', 'Geneva'.

4.2. Imaging Processing

The brain scans were processed in order to have measurements on regions defined in the Desikan-Killiany atlas (Desikan et al., 2006). Brain MRI scans were processed with FreeSurfer [11] (Reuter et al., 2012) to measure brain cortical and sub-cortical volumes, and volumes occupied by the cerebrospinal fluid (CSF), for a total of 99 regions of interest. Relative standardized uptake values (SUVR) were computed for the PET scans (FDG, AV45, AV1451), processed with SPM (Ashburner and Friston, 2000). SUVRs were computed using the cerebellum as reference region, and averaged in the same regions used

[11] surfer.nmr.mgh.harvard.edu

Table 1: Number of subjects per view available in each dataset. The last columns provide the size of the intersection ($\cap$) and union ($\cup$) of subjects with available views. Notice how in the jont set no subject has all the modalities.

View:	clin	MRI	FDG	AV45	AV1451	$\cap$	$\cup$
Dataset							
Adni1	740	730	-	-	-	730	740
Adni2	1324	710	424	417	61	53	1324
Miriad	67	67	-	-	-	67	67
Oasis3	529	489	-	148	-	147	529
Geneva	999	-	65	120	54	15	999
Tot. subjects	3659	1996	489	685	115	0	3659
Tot. datasets	5	4	2	3	2		

Table 2: Mean Squared Error (MSE) of test data from Adni2. All models were trained on all the available datasets by holding-out data from the Adni2 test dataset. 5-folds cross validation of MSE is shown as mean (standard deviation). Best results in boldface are significant with an α level of 0.01 with respect to both competing methods.

View	model		
	DAE	KNN5	ours
clin	0.73 (0.14)	0.44 (0.05)	0.45 (0.07)
MRI	1.23 (0.31)	0.88 (0.15)	**0.70** (0.13)
FDG	4.20 (0.56)	4.15 (0.59)	**1.09** (0.15)
AV45	1.45 (0.35)	1.20 (0.25)	**0.89** (0.15)
AV1451	1.54 (0.82)	1.44 (0.83)	**1.05** (0.45)

for the MRI, except those containing the CSF, for a total of 94 regions of interest.

4.3. Gathering Observations into Views

Observations from the five available datasets (§ 4.1) were grouped into the following views.

1. clin: grouping age and the Mini-Mental cognitive score (MMSE).
2. MRI: grouping brain volumes computed with FreeSurfer.
3. FDG: average brain glucose uptake measured through the analysis of FDG-PET scans.
4. AV45: average brain amyloid deposits measured through the analysis of AV45-PET scans.
5. AV1451: average brain tau protein aggregates measured through the analysis of AV1451-PET scans.

For each subject belonging to the 'Adni1', 'Adni2', 'Miriad' and 'Geneva' datasets, we choose the first available time-point, or baseline. In 'oasis3', since measurements were mostly acquired in different days, we choose to pair nearby time points across modalities into a single one. Time interval between views within one subject was minimal (AV45 vs MRI: $\leq$ 90 days, MRI vs clin: $\leq$ 90 days).

In Tab. 1 we show the number of observations stratified by dataset and view. Size of the intersection ($\cap$) and union ($\cup$) of subjects with available views is also provided. Please note that the only view in common across datasets is the clinical one, composed by MMSE and age features only.

We adjusted all the views feature-wise with *ComBat* (Johnson et al., 2007), a normalization method originally develop in genomics, which has been adopted in neuroimaging studies to reduce unwanted sources of variation in the data due to the differences in acquisition protocols among datasets (Fortin et al., 2017, 2018; Orlhac et al., 2020). In ComBat, we set the variable 'age' as main regressor, and 'Adni2' as reference dataset for the training set. The ComBat reference dataset for testing was the whole training split.

A final feature-wise standardization step was applied by zero centering the data and by rescaling them to have a unity variance. Standardization parameters were computed on the training sets and applied to training and testing sets.

4.4. Experiment 1: Benchmark Validation

The purpose of this experiment is to validate on real data the benchmarked results obtained with the synthetic experiments (§ 3).

As benchmark methods, we choose the best performers on the synthetic experiments, namely KNN5 and DAE. We choose for our model a linear Gaussian parameterization for the likelihood and variational distributions of Eq. (7) and Eq. (8) respectively.

We trained all the models (KNN5, DAE, ours) with data coming from all the datasets except from 'Adni2', left out for testing purposes. We choose the 'Adni2' dataset as testing dataset since it provides all the views, and the highest number of observations per view (Tab. 1).

Prediction performances were evaluated with the Mean Squared Error (MSE) metric, measured on the available views in the testing dataset, reconstructed with Eq. (10). All results were validated by means of 5-folds cross-validation.

Results. In Tab. 2, we show the MSE metric on predicting missing views in the testing dataset with our model and with the benckmark ones. Best results are in boldface, which show a clear advantage of using our model and confirm our findings in the synthetic experiments.

4.5. Experiment 2: Feature Prediction

The purpose of this experiment is to compare, in features prediction experiments, the generalization performance the MCVAE model with respect to our new Multi Task extension (MTMCVE). This experiment was run in three different conditions:

1. Single Task with Internal Benchmark (STIB): when training and testing data are chosen from the same dataset;
2. Single Task with External Benchmark (STEB): when models trained on one dataset are tested on another one;
3. Multi Task Learning (MTL): when models are trained on all the available datasets except the testing one.

In STIB and STEB experiments, both MCVAE and MTMCVAE models are trained on the same views, but while in MCVAE we need to discard observations with missing views from the training set, with MT-MCVAE we can include them

Table 3: Mean Squared Reconstruction Error (the lower the better) measured on test dataset views (clinical scores and imaging derived phenotypes) predicted with the Multi-Channel VAE (MCVAE) and the Multi Task MCVAE (MT-MCVAE). 5-folds cross-validation results shown as average (standard deviation). Models were trained on all the available views in the training dataset, independently of their presence in the testing dataset. Experiments were run in two different conditions: 1) when training and testing data are chosen from the same dataset, or Single Task with Internal Benchmark (STIB) learning case; 2) when models trained on one dataset are tested on another dataset, or Single Task with External Benchmark (STEB) case; In all cases the MT-MCVAE performs either similarly or statistically better than the MCVAE, with alpha levels at 0.05 (*), 0.01 (**), and 0.001 (***).

		view	clin		MRI		FDG		AV45		AV1451	
		model	MCVAE	MT-MCVAE	MCVAE	MT-MCVAE	MCVAE	MT-MCVAE	MCVAE	MT-MCVAE	MCVAE	MT-MCVAE
test dataset	condition	train dataset										
Adni1	STIB	Adni1	0.90 (0.12)	0.89 (0.13)	0.85 (0.11)	0.83 (0.12)*	-	-	-	-	-	-
	STEB	Adni2	0.91 (0.17)	0.77 (0.13)*	1.02 (0.23)	0.85 (0.11)***	-	-	-	-	-	-
		Miriad	0.96 (0.17)	1.14 (0.27)	0.80 (0.14)	0.82 (0.13)*	-	-	-	-	-	-
		Geneva	-	-	-	-	-	-	-	-	-	-
		Oasis3	0.83 (0.30)	0.54 (0.10)*	0.80 (0.15)	0.76 (0.11)*	-	-	-	-	-	-
Adni2	STIB	Adni2	0.83 (0.11)	0.73 (0.15)	0.74 (0.13)	0.70 (0.11)**	0.73 (0.14)	0.59 (0.10)***	1.03 (0.19)	0.80 (0.10)***	1.33 (0.59)	1.18 (0.52)*
	STEB	Adni1	0.77 (0.18)	0.80 (0.14)	0.74 (0.11)	0.75 (0.12)	-	-	-	-	-	-
		Miriad	0.73 (0.20)	0.71 (0.18)	0.78 (0.13)	0.77 (0.13)	-	-	-	-	-	-
		Geneva	0.47 (0.06)	0.48 (0.09)	-	-	1.40 (0.21)	1.09 (0.15)***	1.10 (0.21)	0.91 (0.15)**	1.34 (0.52)	1.05 (0.45)***
		Oasis3	0.76 (0.23)	0.61 (0.13)	0.68 (0.12)	0.68 (0.11)	-	-	1.32 (0.29)	1.13 (0.26)***	-	-
Gneva	STIB	Geneva	0.79 (0.34)	0.98 (0.52)	-	-	3.63 (1.35)	3.18 (1.04)*	1.82 (0.57)	1.76 (0.47)*	1.27 (0.82)	1.19 (0.67)*
	STEB	Adni1	-	-	-	-	-	-	-	-	-	-
		Adni2	2.57 (1.09)	2.07 (1.05)	-	-	3.01 (1.05)	2.69 (0.77)*	1.92 (0.90)	1.41 (0.39)	1.81 (0.81)	1.42 (0.66)***
		Miriad	-	-	-	-	-	-	-	-	-	-
		Oasis3	1.93 (0.66)	2.28 (0.89)	-	-	-	-	1.70 (0.51)	1.63 (0.55)*	-	-
Miriad	STIB	Miriad	3.21 (1.07)	3.23 (2.55)	6.39 (1.57)	6.38 (1.52)	-	-	-	-	-	-
	STEB	Adni1	6.90 (3.33)	6.49 (3.42)	6.60 (1.61)	6.73 (1.55)	-	-	-	-	-	-
		Adni2	5.60 (2.76)	3.97 (3.14)	5.93 (1.90)	6.59 (1.64)	-	-	-	-	-	-
		Geneva	-	-	-	-	-	-	-	-	-	-
		Oasis3	6.80 (6.52)	6.24 (4.62)	6.29 (1.68)	6.23 (1.40)	-	-	-	-	-	-
Oasis3	STIB	Oasis3	0.83 (0.33)	0.68 (0.28)	0.68 (0.13)	0.66 (0.12)*	-	-	1.58 (0.63)	1.22 (0.26)***	-	-
	STEB	Adni1	1.20 (0.25)	1.23 (0.28)	0.78 (0.14)	0.79 (0.14)	-	-	-	-	-	-
		Adni2	1.11 (0.33)	1.09 (0.24)	0.89 (0.18)	0.76 (0.15)***	-	-	0.94 (0.22)	1.02 (0.26)*	-	-
		Miriad	0.98 (0.21)	1.02 (0.20)	0.83 (0.18)	0.83 (0.18)	-	-	-	-	-	-
		Geneva	0.55 (0.28)	0.49 (0.26)	-	-	-	-	1.23 (0.61)	1.11 (0.26)*	-	-

Table 4: Mean Squared Reconstruction Error (the lower the better) measured on test dataset views (clinical scores and imaging derived phenotypes) predicted with our model. 5-folds cross-validation results shown as as average (standard deviation). Models were trained on all the available views in the training dataset, independently of their presence in the testing dataset. Experiments were run in three different conditions: 1) when training and testing data are chosen from the same dataset, or Single Task with Internal Benchmark (STIB) learning case; 2) when models trained on one dataset are tested on another dataset, or Single Task with External Benchmark (STEB) case; 3) when models are trained on all the available datasets except the testing one, or Multi Task Learning (MTL). We measure a better performance in the MTL condition with respect to the STIB (§) in $7/12$ of cases, and in $10/12$ of cases with respect to the average STEB (†) experiments.

	view	clin	MRI	AV45
test dataset	condition			
Adni1	STIB	0.89 (0.13)	0.83 (0.12)	-
	STEB (avg)	0.82 (0.17)	0.81 (0.12)	-
	MTL	0.45 (0.07)§†	0.77 (0.10)§†	-
Adni2	STIB	0.73 (0.15)	0.70 (0.11)	0.80 (0.10)
	STEB (avg)	0.65 (0.14)	0.73 (0.12)	1.02 (0.21)
	MTL	0.45 (0.07)§†	0.70 (0.13)†	0.89 (0.15)†
Geneva	STIB	0.98 (0.52)	-	1.76 (0.47)
	STEB (avg)	2.18 (0.97)	-	1.52 (0.47)
	MTL	1.80 (1.16)†	-	1.35 (0.37)§†
Miriad	STIB	3.23 (2.55)	6.38 (1.52)	-
	STEB (avg)	5.57 (3.73)	6.52 (1.53)	-
	MTL	2.31 (1.65)§†	6.17 (1.37)§†	-
Oasis3	STIB	0.68 (0.28)	0.66 (0.12)	1.22 (0.26)
	STEB (avg)	0.96 (0.25)	0.79 (0.16)	1.07 (0.26)
	MTL	0.72 (0.09)†	0.81 (0.15)	1.09 (0.30)§

by grouping together observations with common views into homogeneous tasks. In MTL experiments, MCVAE models cannot be trained because no observation has simultaneously all the views.

We choose for both MCVAE and MT-MCVAE a linear Gaussian parameterization for the likelihood and variational distributions as in Eq. (7) and Eq. (8) respectively. Models were trained on all the available views in the training dataset. Prediction performances were evaluated with the Mean Squared Error (MSE) metric, measured on the available views in the testing dataset, reconstructed with Eq. (10). All results were validated by means of 5-folds cross-validation.

Results. In Tab. 3 and Tab. 4 we show the prediction error in terms of MSE for each test dataset and view, on the three experimental conditions described earlier. In STIB and STEB cases (Tab. 3), the MT-MCVAE model performs either similarly or statistically better than the MCVAE, especially in cases where the difference between the union and intersection set of observations is higher (cfr. Tab. 1). In the MTL scenario (Tab. 4) there are 12 cases that could be fitted with MT-MCVAE only. We measure an overall better performance of MTL with respect to STIB ($7/12$ of cases) and with respect to STEB ($10/12$ of cases).

4.6. Experiment 3: Diagnosis Prediction

The purpose of this experiment is to compare, in diagnosis prediction experiments, the generalization performance of the

Table 5: Number of subjects stratified by dataset and diagnosis: Alzheimer's Disease (AD); Mild Cognitive Impairment (MCI); Normal Cognition (NC).

	AD	MCI	NC	other	total
Adni1	403	172	165	-	740
Adni2	328	455	541	-	1324
Geneva	147	405	-	447	999
Miriad	44	-	23	-	67
Oasis3	149	-	380	-	529

MCVAE model with respect to the MT-MCVAE, in the three experimental conditions described earlier: STIB, STEB, and MTL. Diagnostic classes are: Alzheimer's disease (AD), mild cognitive impairment (MCI), normal cognition (NC).

For both MCVAE and MT-MCVAE we choose a linear Gaussian parameterization for the variational distributions as in Eq. (8). To adapt the models to this new classification experiment, we adopt as decoding function for the latent variable $\mathbf{z}$, the following Categorical likelihood:

$$p\left(y_{d,n}|\mathbf{z},\boldsymbol{\theta}\right) = \mathrm{Cat}\left(\pi = \boldsymbol{\theta}\mathbf{z}\right), \tag{12}$$

where $y_{d,n}$ is the diagnosis associated to the data-point n in the dataset d. The probability vector π is a two dimensional vector representing the class probability for each of the three binary comparisons across the three diagnostic classes, namely AD *vs* MCI, AD *vs* NC, MCI *vs* NC ,and is parametrized with a linear transformation of the latent $\mathbf{z}$ by the matrix $\boldsymbol{\theta}$.

Models were trained on all the available views in the training dataset, independently of their presence in the testing dataset. Classes probabilities were inferred from the all the available views in the testing dataset with the following equation:

$$\hat{y}_{d,n} = \frac{1}{V_{d,n}} \sum_{w=1}^{V_{d,n}} \mathbb{E}_{q_{d,n,w}(\mathbf{z})}\left[p\left(y_{d,n}|\mathbf{z},\boldsymbol{\theta}\right)\right]. \tag{13}$$

We attributed to each subject the diagnostic class with the highest inferred probability.

The performance on test datasets was evaluated by measuring the classification accuracy (%). All results were validated by means of 5-folds cross-validation.

Results. In Tab. 6 we show the classification accuracy of MCVAE and MT-MCVAE. In STIB and STEB cases, the MT-MCVAE model performs either similarly or statistically better than the MCVAE. There are 7 cases in the MTL condition (Tab. 7) that could be fitted with the MT-MCVAE model only. In all of them we measure a better performance with respect to the best STEB cases.

5. Discussion

In both the experiments on synthetic and real data, our model compared favorably with respect to state of the art benchmark methods.

An interesting result is the one presented in Fig. 4, suggesting that collecting a minimum amount of data-points with complete views (25% in our setup) is enough for our model to capture the joint relationship among views. In our synthetic tests, training on more complete scenarios does not seem to improve significantly the testing results. This condition may be explained by the high collinearity between features due to the linear mappings used to generate the multi-view data.

As a secondary result, we report the positive performance of knn (k = 5) in synthetic scenarios, especially in low snr cases, and on real data experiment, were it is most of the time superior to the DAE. This finding is corroborated by Platias and Petasis (2020) were knn is found to be superior to methods based on autoencoders.

The experimental results on real medical imaging datasets (Tab. 3, Tab. 6) show the clear improvement of our method (MT-MCVAE) with respect to the MCVAE, that inspired our work. The features and diagnosis prediction clearly improves when using our method, that allows to not discard observations with missing views.

In feature prediction experiments (Tab. 4) we showed that MT-MCVAE models trained jointly on multiple datasets (MTL cases) perform generally better than the ones trained on a single dataset. We suspect that there are two reasons explaining these results. The first is that modeling simultaneously multiple datasets with our method brings more variability and information at play, making the generalization to unseen data less prone to prediction errors. The second reason maybe that every decoder, associated to its specific view, acts, through the shared latent space, as a regularizer for all the other decoders.

In diagnosis prediction experiments (Tab. 7) the MT-MCVAE generalizes better to new unseen datasets when trained jointly on multiple datasets (MTL cases) with respect to cases where the training happens on a single dataset. We notice that the best results happen in cases where testing data and training data come from the same dataset (ST cases), that is when the testing dataset is not anymore unseen to our model. This is a different result than the analogous one in the feature prediction experiments, and we argue that the reason may be due to the lack of the regularization mechanism induced by having concurring decoders. Indeed, the MT-MCVAE classifier is composed by a single decoder only, which can become highly specialized in decoding testing data coming from the same dataset of the training data.

In our work we have thoroughly investigated architectures with a one-to-one correspondence between encoding and decoding views. This makes our model part of the family of the auto-encoders, where the model acts as identity transformation between the input and the output. Other architectures are nevertheless possible, such as the classifier described in § 4.6 In general, there may be an m-to-n relationship, with partially overlapping views among m input views and n output views. Investigating the properties of all the possible architectures is beyond the scope of this work.

As final remark, we want to stress that our model is based on the assumption of independent and identical distributed observations. This assumption may be limiting in healthcare datasets, such as the ones used in this work. In our work we mitigated these biases by harmonizing the datasets before ap-

Table 6: Experiment of diagnosis classification run with the Multi-Channel VAE (MCVAE) and the Multi Task MCVAE (MT-MCVAE). 5-folds classification accuracy in % is shown as mean (standard deviation). Since there are no MCI in miriad and oasis3 datasets, the classification tests 'AD vs MCI' and 'MCI vs NC' are meaningless and not reported. Since there are no NC in the geneva dataset, the classification tests 'AD vs NC' and 'MCI vs NC' are meaningless and not reported. Experiments were run in two different conditions: 1) when training and testing data are chosen from the same dataset, or Single Task with Internal Benchmark (STIB) learning case; 2) when models trained on one dataset are tested on another dataset, or Single Task with External Benchmark (STEB) case. In all cases the MT-MCVAE model performs either similarly or statistically better than the MCVAE, with alpha levels at 0.05 (*), 0.01 (**), and 0.001 (***).

		classification task	AD vs MCI		AD vs NC		MCI vs NC	
		model	MCVAE	MT-MCVAE	MCVAE	MT-MCVAE	MCVAE	MT-MCVAE
test dataset	condition	train dataset						
Adni1	STIB	Adni1	72.70 (3.72)	72.87 (4.37)	81.69 (2.97)	81.51 (3.14)	62.00 (8.91)	62.90 (8.72)
	STEB	Adni2	47.48 (3.56)	58.96 (3.55)***	68.50 (4.86)	73.77 (2.80)*	53.12 (6.42)	59.65 (2.76)*
		Miriad	-	-	82.58 (4.75)	80.82 (3.16)	-	-
		Oasis3	-	-	48.57 (6.48)	62.31 (6.43)**	-	-
		Geneva	36.52 (5.29)	46.61 (8.03)*	-	-	-	-
Adni2	STIB	Adni2	50.58 (3.90)	80.07 (2.53)***	82.86 (3.28)	87.92 (3.46)*	58.63 (4.27)	65.56 (1.11)**
	STEB	Adni1	57.59 (2.61)	58.23 (2.87)	64.21 (3.36)	64.21 (3.52)	63.05 (2.00)	62.75 (1.80)
		Miriad	-	-	70.32 (7.29)	70.20 (7.17)	-	-
		Oasis3	-	-	68.24 (2.97)	75.72 (1.90)**	-	-
		Geneva	64.49 (2.98)	63.98 (3.30)	-	-	-	-
Geneva	STIB	Geneva	65.76 (3.62)	77.70 (8.12)*	-	-	-	-
	STEB	Adni1	29.17 (5.87)	30.08 (5.49)	-	-	-	-
		Adni2	38.61 (15.08)	70.11 (2.90)**	-	-	-	-
Miriad	STIB	Miriad	-	-	83.85 (13.84)	86.70 (15.68)	-	-
	STEB	Adni1	-	-	74.18 (14.37)	74.18 (14.37)	-	-
		Adni2	-	-	74.95 (11.58)	78.90 (11.54)*	-	-
		Oasis3	-	-	45.71 (18.08)	66.04 (19.35)	-	-
Oasis3	STIB	Oasis3	-	-	74.47 (2.49)	80.35 (3.59)*	-	-
	STEB	Adni1	-	-	49.16 (6.34)	48.22 (5.78)	-	-
		Adni2	-	-	67.86 (3.80)	75.42 (4.68)*	-	-
		Miriad	-	-	64.48 (8.65)	62.02 (9.74)	-	-

Table 7: Experiment of diagnosis classification run with our model. 5-folds classification accuracy in % is shown as mean (standard deviation). Experiments were run in three different conditions: 1) when training and testing data are chosen from the same dataset, or Single Task with Internal Benchmark (STIB) learning case; 2) when models trained on one dataset are tested on another dataset, or Single Task with External Benchmark (STEB) case; 3) when models are trained on all the available datasets except the testing one, or Multi Task Learning (MTL). In all cases we measure a better performance in the MTL condition with respect to the average STEB one (†).

	classification task	AD vs MCI	AD vs NC
test dataset	condition		
Adni1	STIB	72.87 (4.37)	81.51 (3.14)
	STEB	52.79 (5.79)	72.30 (4.13)
	MTL	56.17 (4.29)†	74.48 (3.06)†
Adni2	STIB	80.07 (2.53)	87.92 (3.46)
	STEB	61.11 (3.09)	70.04 (4.20)
	MTL	68.46 (1.31)†	83.89 (1.93)†
Geneva	STIB	77.70 (8.12)	-
	STEB	50.10 (4.20)	-
	MTL	56.70 (4.13)†	-
Miriad	STIB	-	86.70 (15.68)
	STEB	-	73.04 (15.09)
	MTL	-	92.53 (0.30)†
Oasis3	STIB	-	80.35 (3.59)
	STEB	-	61.89 (6.73)
	MTL	-	75.43 (5.67)†

plying our model, and we leave the extension and development of a bias-transparent multi-view models to future works.

6. Conclusions

We proposed a new multi-task latent variable generative model able to learn simultaneously from multiple datasets, even in the presence of non-overlapping views among all the datasets. The available overlap between pairs of datasets allows the information to flow through all the views in the dataset pool. Since the learned view-specific parameters are shared among datasets, missing views can be automatically imputed for every dataset. The method proposed in this work is a coherent extension of classical variational generative models, making the training fast and scalable. Being dataset agnostic, our method allows to integrate all the available data into a joint model, gathering all the available information from multiple datasets at the same time. We conducted extensive tests for the joint modeling of synthetically generated data and of multi-modal neuroimaging datasets from independent dementia studies and associated clinical data, showing the competitiveness of our method with respect to the state of the art. Thanks to its general formulation, the proposed method can find applications beyond the neuroimaging research field.

Declaration of interests

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.

Acknowledgments

This work has been supported by:

- the French government, through the UCA$^{\text{JEDI}}$ Investments in the Future project managed by the National Research Agency (ANR) with the reference number ANR-15-IDEX-01;
- the French government, through the 3IA Côte d'Azur Investments in the Future project managed by the National Research Agency (ANR) with the reference number ANR-19-P3IA-0002;
- the OPAL infrastructure from Université Côte d'Azur, providing computational resources and support;
- the "Centre de la Mémoire" at Geneva University Hospitals, funded by private donors: A.P.R.A. - Association Suisse pour la Recherche sur la Maladie d'Alzheimer, Genève; Fondation Segré, Genève; Ivan Pictet, Genève; Fondazione Agusta, Lugano; Fondation Chmielewski, Genève. Competitive research projects have been funded by: H2020, Human Brain Project, Innovative Medicines Initiative (IMI), IMI2, Swiss National Science Foundation (project numbers: 169876, 185028, 182772); VELUX Foundation (project number 1123).
- 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 NeuroImaging at the University of Southern California.

- Data used in the preparation of this article were obtained also from the MIRIAD database. The MIRIAD investigators did not participate in analysis or writing of this report. The MIRIAD dataset is made available through the support of the UK Alzheimer's Society (Grant RF116). The original data collection was funded through an unrestricted educational grant from GlaxoSmithKline (Grant 6GKC).
- Data were also provided by OASIS-3: Principal Investigators: T. Benzinger, D. Marcus, J. Morris; NIH P50AG00561, P30NS09857781, P01AG026276, P01AG003991, R01AG043434, UL1TR000448, R01EB009352. AV-45 doses were provided by Avid Radiopharmaceuticals, a wholly owned subsidiary of Eli Lilly. Avid has no involvement in study design or analysis.

References

Antelmi, L., Ayache, N., Robert, P., Lorenzi, M., 2019. Sparse Multi-Channel Variational Autoencoder for the Joint Analysis of Heterogeneous Data, in: Chaudhuri, K., Salakhutdinov, R. (Eds.), Proc. 36th Int. Conf. Mach. Learn., PMLR. pp. 302–311.

Ashburner, J., Friston, K.J., 2000. Voxel-based morphometry–the methods. Neuroimage 11, 805–21.

Blei, D.M., Kucukelbir, A., McAuliffe, J.D., 2017. Variational Inference: A Review for Statisticians. J. Am. Stat. Assoc. 112, 859–877. doi:10.1080/01621459.2017.1285773.

Buch, A.M., Liston, C., 2020. Dissecting diagnostic heterogeneity in depression by integrating neuroimaging and genetics. Neuropsychopharmacology doi:10.1038/s41386-020-00789-3.

van Buuren, S., Groothuis-Oudshoorn, C.G.M., 2000. Multivariate Imputation by Chained Equations: MICE V1.0 User manual. volume PG/VGZ/00.038. TNO Prevention and Health, Leiden.

Caruana, R., 1998. Multitask Learning. Springer US, Boston, MA. pp. 95–133. doi:10.1007/978-1-4615-5529-2_5.

Chen, T., Ma, X., Liu, X., Wang, W., Feng, R., Chen, J., Yuan, C., Lu, W., Chen, D.Z., Wu, J., 2019. Multi-view Learning with Feature Level Fusion for Cervical Dysplasia Diagnosis, pp. 329–338. doi:10.1007/978-3-030-32239-7_37.

Desikan, R.S., Ségonne, F., Fischl, B., Quinn, B.T., Dickerson, B.C., Blacker, D., Buckner, R.L., Dale, A.M., Maguire, R.P., Hyman, B.T., Albert, M.S., Killiany, R.J., 2006. An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. Neuroimage doi:10.1016/j.neuroimage.2006.01.021.

Dorado-Moreno, M., Navarin, N., Gutiérrez, P., Prieto, L., Sperduti, A., Salcedo-Sanz, S., Hervás-Martínez, C., 2020. Multi-task learning for the prediction of wind power ramp events with deep neural networks. Neural Networks 123, 401–411. doi:10.1016/j.neunet.2019.12.017.

Finn, C., Abbeel, P., Levine, S., 2017. Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks. PMLR 70, 1126–1135.

Fortin, J.P., Cullen, N., Sheline, Y.I., Taylor, W.D., Aselcioglu, I., Cook, P.A., Adams, P., Cooper, C., Fava, M., McGrath, P.J., McInnis, M., Phillips, M.L., Trivedi, M.H., Weissman, M.M., Shinohara, R.T., 2018. Harmonization of cortical thickness measurements across scanners and sites. Neuroimage 167, 104–120.

Fortin, J.P., Parker, D., Tunç, B., Watanabe, T., Elliott, M.A., Ruparel, K., Roalf, D.R., Satterthwaite, T.D., Gur, R.C., Gur, R.E., Schultz, R.T., Verma, R., Shinohara, R.T., 2017. Harmonization of multi-site diffusion tensor imaging data. Neuroimage 161, 149–170. doi:10.1016/j.neuroimage.2017.08.047.

Golriz Khatami, S., Robinson, C., Birkenbihl, C., Domingo-Fernández, D., Hoyt, C.T., Hofmann-Apitius, M., 2020. Challenges of Integrative Disease Modeling in Alzheimer's Disease. Front. Mol. Biosci. 6. doi:10.3389/fmolb.2019.00158.

Gondara, L., Wang, K., 2018. Mida: Multiple imputation using denoising autoencoders, in: Phung, D., Tseng, V.S., Webb, G.I., Ho, B., Ganji, M., Rashidi, L. (Eds.), Advances in Knowledge Discovery and Data Mining, Springer International Publishing, Cham. pp. 260–272.

Johnson, W.E., Li, C., Rabinovic, A., 2007. Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics 8, 118–127. doi:10.1093/biostatistics/kxj037.

Kingma, D.P., Salimans, T., Welling, M., 2015. Variational Dropout and the Local Reparameterization Trick, in: Cortes, C., Lawrence, N.D., Lee, D.D., Sugiyama, M., Garnett, R. (Eds.), Adv. Neural Inf. Process. Syst. 28. Curran Associates, Inc., pp. 2575–2583.

Kingma, D.P., Welling, M., 2014. Auto-Encoding Variational Bayes, in: Proc. 2nd Int. Conf. Learn. Represent. (ICLR2014).

LaMontagne, P.J., Benzinger, T.L., Morris, J.C., Keefe, S., Hornbeck, R., Xiong, C., Grant, E., Hassenstab, J., Moulder, K., Vlassenko, A., Raichle, M.E., Cruchaga, C., Marcus, D., 2019. Oasis-3: Longitudinal neuroimaging, clinical, and cognitive dataset for normal aging and alzheimer disease. medRxiv doi:10.1101/2019.12.13.19014902.

Le Sueur, H., Bruce, I.N., Geifman, N., 2020. The challenges in data integration – heterogeneity and complexity in clinical trials and patient registries of Systemic Lupus Erythematosus. BMC Med. Res. Methodol. 20, 164. doi:10.1186/s12874-020-01057-0.

Malone, I.B., Cash, D., Ridgway, G.R., MacManus, D.G., Ourselin, S., Fox, N.C., Schott, J.M., 2013. MIRIAD—Public release of a multiple time point Alzheimer's MR imaging dataset. Neuroimage 70, 33–36. doi:10.1016/j.neuroimage.2012.12.044.

Molchanov, D., Ashukha, A., Vetrov, D., 2017. Variational Dropout Sparsifies Deep Neural Networks. arXiv .

Orlhac, F., Lecler, A., Savatovski, J., Goya-Outi, J., Nioche, C., Charbonneau, F., Ayache, N., Frouin, F., Duron, L., Buvat, I., 2020. How can we combat multicenter variability in MR radiomics? Validation of a correction procedure. Eur. Radiol. doi:10.1007/s00330-020-07284-9.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., Chintala, S., 2019. Pytorch: An imperative style, high-performance deep learning library, in: Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché Buc, F., Fox, E., Garnett, R. (Eds.), Advances in Neural Information Processing Systems 32. Curran Associates, Inc., pp. 8024–8035.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E., 2011. Scikit-learn: Machine learning in Python. Journal of Machine Learning Research 12, 2825–2830.

Platias, C., Petasis, G., 2020. A Comparison of Machine Learning Methods for Data Imputation, in: 11th Hell. Conf. Artif. Intell., ACM. pp. 150–159. doi:10.1145/3411408.3411465.

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

Rezende, D.J., Mohamed, S., Wierstra, D., 2014. Stochastic Backpropagation and Approximate Inference in Deep Generative Models .

Shi, Y., N, S., Paige, B., Torr, P., 2019. Variational Mixture-of-Experts Autoencoders for Multi-Modal Deep Generative Models, in: Wallach, H., Larochelle, H., Beygelzimer, A., d\textquotesingle Alché-Buc, F., Fox, E., Garnett, R. (Eds.), Adv. Neural Inf. Process. Syst. 32. Curran Associates, Inc., pp. 15718–15729.

Tognin, S., van Hell, H.H., Merritt, K., Winter-van Rossum, I., Bossong, M.G., Kempton, M.J., Modinos, G., Fusar-Poli, P., Mechelli, A., Dazzan, P., Maat, A., de Haan, L., Crespo-Facorro, B., Glenthøj, B., Lawrie, S.M., McDonald, C., Gruber, O., van Amelsvoort, T., Arango, C., Kircher, T., Nelson, B., Galderisi, S., Bressan, R., Kwon, J.S., Weiser, M., Mizrahi, R., Sachs, G., Maatz, A., Kahn, R., McGuire, P., McGuire, P., Tognin, S., Fusar-Poli, P., Kempton, M., Modinos, G., Merritt, K., Mechelli, A., Dazzan, P., Gifford, G., Petros, N., Antoniades, M., De Micheli, A., Vieira, S., Spencer, T.J., Scarpazza, C., Hird, E., Kahn, R., Maat, A., van Hell, E., Winter, I., Cahn, W., Schnack, H., de Haan, L., Siegmann, D., Barkhof, J., Hendriks, L., de Wit, I., Crespo-Facorro, B., Tordesillas-Gutierrez, D., Setien-Suero, E., Ayesa-Arriola, R., Suarez-Pinilla, P., Ramirez-Bonilla, M., Garcia-de la foz, V.O., Glenthøj, B., Erlang Sørensen, M., Tangmose, K., Schæbel, H., Broberg, B., Rostrup, E., Lawrie, S., McDonald, C., Hallahan, B., Cannon, D., McLoughlin, J., Finnegan, M., Gruber, O., van Amelsvoort, T., Deckers, D., Marcelis, M., Vingerhoets, C., Arango, C., Díaz-Caneja, C.M., Ayora, M., Janssen, J., Rodríguez-Jiménez, R., Díaz-Marsá, M., Kircher, T., Falkenberg, I., Bitsch, F., Berger, P., Sommer, J.,

Raab, K., Jakobi, B., Nelson, B., McGorry, P., Amminger, P., McHugh, M., Galderisi, S., Mucci, A., Bucci, P., Piegari, G., Pietrafesa, D., Nicita, A., Patriarca, S., Bressan, R., Zugman, A., Gadelha, A., Rodrigues da Cunha, G., Soo Kwon, J., Kevin Cho, K.I.k., Young Lee, T., Kim, M., Bin Kwak, Y., Jeong Hwang, W., Weiser, M., Mizrahi, R., Kiang, M., Gerritsen, C., Maheandiran, M., Ahmed, S., Prce, I., Lepock, J., Sachs, G., Willeit, M., Lenczowski, M., Sauerzopf, U., Weidenauer, A., Furtner-Srajer, J., Kirschner, M., Maatz, A., Burrer, A., Stämpfli, P., Huber, N., Kaiser, S., Kawohl, W., Brammer, M., Young, J., Bullmore, E., Morgan, S., 2020. Towards Precision Medicine in Psychosis: Benefits and Challenges of Multimodal Multicenter Studies—PSYSCAN: Translating Neuroimaging Findings From Research into Clinical Practice. Schizophr. Bull. 46, 432–441. doi:10.1093/schbul/sbz067.

van Buuren, S., Groothuis-Oudshoorn, K., 2011. mice: Multivariate imputation by chained equations in r. Journal of Statistical Software 45, 1–67. URL: https://www.jstatsoft.org/v45/i03/.

Vieira, S., Pinaya, W.H.L., Garcia-Dias, R., Mechelli, A., 2020. Multimodal integration, in: Mach. Learn. - Methods Appl. to Brain Disord.. Elsevier. chapter 16, pp. 283–305. doi:10.1016/B978-0-12-815739-8.00016-X.

Vincent, P., Larochelle, H., Bengio, Y., Manzagol, P.A., 2008. Extracting and composing robust features with denoising autoencoders, in: Proc. 25th Int. Conf. Mach. Learn. - ICML '08, ACM Press, New York, New York, USA. pp. 1096–1103. doi:10.1145/1390156.1390294.

Wang, S., Manning, C., 2013. Fast dropout training. Proc. 30th Int. Conf. Mach. Learn. 28, 118–126.

Wei, W., Poirion, E., Bodini, B., Durrleman, S., Ayache, N., Stankoff, B., Colliot, O., 2019. Predicting PET-derived demyelination from multimodal MRI using sketcher-refiner adversarial training for multiple sclerosis. Med. Image Anal. 58. doi:10.1016/j.media.2019.101546.

Wu, M., Goodman, N., 2018. Multimodal Generative Models for Scalable Weakly-Supervised Learning, in: Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., Garnett, R. (Eds.), Adv. Neural Inf. Process. Syst. 31. Curran Associates, Inc., pp. 5575–5585.

Zhou, T., Fu, H., Chen, G., Shen, J., Shao, L., 2020. Hi-Net: Hybrid-Fusion Network for Multi-Modal MR Image Synthesis. IEEE Trans. Med. Imaging 39, 2772–2781. doi:10.1109/TMI.2020.2975344.