Disease progression score estimation from multimodal imaging and microRNA data using supervised variational autoencoders

Virgilio Kmetzsch, Emmanuelle Becker, Dario Saracino, Daisy Rinaldi, Agnes Camuzat, Isabelle Le Ber, Olivier Colliot

Disease progression score estimation from multimodal imaging and microRNA data using supervised variational autoencoders

Virgilio Kmetzsch, Emmanuelle Becker, Dario Saracino, Daisy Rinaldi, Agnès Camuzat, Isabelle Le Ber, and Olivier Colliot, *Member, IEEE*, for the PREV-DEMALS study group

***Abstract*— Frontotemporal dementia and amyotrophic lateral sclerosis are rare neurodegenerative diseases with no effective treatment. The development of biomarkers allowing an accurate assessment of disease progression is crucial for evaluating new therapies. Concretely, neuroimaging and transcriptomic (microRNA) data have been shown useful in tracking their progression. However, no single biomarker can accurately measure progression in these complex diseases. Additionally, large samples are not available for such rare disorders. It is thus essential to develop methods that can model disease progression by combining multiple biomarkers from small samples. In this paper, we propose a new framework for computing a disease progression score (DPS) from cross-sectional multimodal data. Specifically, we introduce a supervised multimodal variational autoencoder that can infer a meaningful latent space, where latent representations are placed along a disease trajectory. A score is computed by orthogonal projections onto this path. We evaluate our framework with multiple synthetic datasets and with a real dataset containing 14 patients, 40 presymptomatic genetic mutation carriers and 37 controls from the PREV-DEMALS study. There is no ground truth for the DPS in real-world scenarios, therefore we use the area under the ROC curve (AUC) as a proxy metric. Results with the synthetic datasets support this choice, since the higher the AUC, the more accurate the predicted simulated DPS. Experiments with the real dataset demonstrate better performance in comparison with state-of-the-art approaches. The proposed framework thus leverages cross-sectional multimodal datasets with small sample sizes to objectively measure disease progression, with potential application in clinical trials.**

The research leading to these results has received funding from the French government under management of Agence Nationale de la Recherche, references ANR-19-P3IA-0001 (PRAIRIE 3IA Institute), ANR-10-IAIHU-06, project PREV-DEMALS (grant number ANR-14-CE15-0016-07), and from the Inria Project Lab Program (project Neuromarkers).

V. Kmetzsch, D. Saracino and O. Colliot are with Sorbonne Université, Institut du Cerveau - Paris Brain Institute - ICM, CNRS, Inria, Inserm, AP-HP, Paris, France (e-mails: virgilio.kmetzsch@inria.fr, dario.saracino@icm-institute.org, olivier.colliot@cnrs.fr).

E. Becker is with Univ Rennes, Inria, CNRS, IRISA, F-35000, Rennes, France (e-mail: emmanuelle.becker@univ-rennes1.fr).

D. Rinaldi and A. Camuzat are with Sorbonne Université, Institut du Cerveau - Paris Brain Institute - ICM, CNRS, Inserm, AP-HP, Paris, France (daisy.rinaldi@icm-institute.org, agnes.camuzat@icm-institute.org).

I. Le Ber is with Sorbonne Université, Institut du Cerveau - Paris Brain Institute - ICM, CNRS, Inserm, AP-HP, Hôpital de la Pitié Salpêtrière, Institute of Memory and Alzheimer's Disease (IM2A) , F-75013, Paris, France (e-mail: isabelle.leber@icm-institute.org).

***Index Terms*— Disease progression score, Deep learning, MicroRNA, Multimodal data, Neurodegenerative disease, Neuroimaging, Variational autoencoder**

I. Introduction

FRONTOTEMPORAL dementia (FTD) and amyotrophic lateral sclerosis (ALS) are rare neurodegenerative disorders that have devastating personal and social consequences. FTD and ALS may be sporadic (no previous family history) or genetically inherited. The most common genetic cause of FTD and ALS is a hexanucleotide repeat expansion in the *C9orf72* gene [1], [2]. These fatal conditions can sometimes coexist in *C9orf72*-mutated individuals, and have no cure or standard treatment to date.

Carriers of the *C9orf72* mutation that do not present clinical symptoms are considered presymptomatic, since they have a very high probability of manifesting FTD and/or ALS later in life. Clinical trials for potential therapies are likely to be most effective at this presymptomatic stage, before any irreversible brain damage has occurred. However, the evaluation of new treatments depends on an accurate measure of disease progression, which is not evident without observable symptoms. Previous work has shown the relevance of neuroimaging [3], [4] and transcriptomic (microRNA) [5] biomarkers for a better understanding of *C9orf72*-disease in presymptomatic carriers. Nevertheless, when these modalities are analysed separately, they provide only an incomplete picture of these diseases. It is thus essential to develop methods that combine different modalities to accurately measure disease progression. As different biomarkers characterise distinct disease stages, various biomarkers can be combined to represent the entire disease course with a single measure, commonly referred in the literature as the *disease progression score* (DPS).

The idea of computing disease progression scores falls within the larger topic of modeling disease progression. In the past years, many approaches have been developed for data-driven modeling of disease progression, such as event-based models (EBM) [6]–[9], different algorithms fitting logistic functions to biomarker trajectories [10], [11], non-linear mixed-effects models [12], [13], a vertex-wise model of brain diseases fitted with expectation-maximisation [14], Gaussian processes [15]–[17], topological profiles reflecting

brain connectivity [18], Bayesian multi-task learning [19], and recurrent neural networks [20].

Most of these approaches require longitudinal data. For instance, the authors of [10] assume that the longitudinal dynamic of each biomarker can be represented as a sigmoidal function of the DPS. They propose a joint optimization algorithm to compute the DPS, fit one sigmoid function per biomarker using alternating least squares, and apply their work to hundreds of patients with Alzheimer's disease (AD). Similarly, a more recent method [11], also applied to AD, uses M-estimation to map each subject's age to a DPS, jointly fitting generalized logistic functions to the longitudinal dynamics of biomarkers as functions of the DPS. Schiratti et al [12] proposed a general non-linear mixed-effects model for longitudinal data based on concepts from Riemannian geometry. The application of this framework to AD, called AD Course Map [13], allowed to map each subject to their corresponding disease stage. The authors of [15] proposed a probabilistic approach based on Gaussian process regression from time-series of biomarker measurements. Yet another framework, named Data-driven Inference of Vertexwise Evolution (DIVE) [14] consists in identifying clusters of vertexwise biomarker measurements in the brain, and estimating representative trajectories for these clusters. Finally, [20] uses recurrent neural networks to predict biomarker values without parametric assumptions about trajectories, with application to AD. To the best of our knowledge, the only disease modeling approaches that infer a DPS from cross-sectional data are EBM [6]–[9]. These models explore the temporal sequence in which biomarkers become abnormal in the course of a disease. They have been successfully applied to a variety of diseases including AD [6]–[9], [21]–[23], multiple sclerosis [24], [25], Parkinson's disease [26], Huntington's disease [27] as well as FTD [28], [29] and ALS [30]. However, in these works, EBMs were applied to a relatively small number of features (typically 10-50). Although these are the state-of-the-art methods for disease progression modeling with cross-sectional datasets, previous studies do not clarify if they would perform well in higher dimensions.

Despite the recognized importance of estimating neurodegenerative diseases progression, research has tended to focus mostly on higher prevalence conditions. Existing solutions are thus inadequate to model rare diseases with high-dimensional cross-sectional data, for three main reasons. First, we observe that longitudinal data is needed for the vast majority of approaches. However, *C9orf72*-associated FTD and ALS are slowly progressive conditions in the presymptomatic phase, which hinders the collection of meaningful longitudinal data. Second, most published methods benefit from large samples, which are not available for very low prevalence disorders such as genetic FTD and ALS. Finally, it is unclear if event-based models, the only methods suitable for cross-sectional data, can be robustly applied to high-dimensional microRNA expression data, which comprise hundreds of biomarkers.

In this paper, we present a novel framework to estimate disease progression scores for rare neurodegenerative disorders using only cross-sectional multimodal data. To that purpose, we introduce a new supervised multimodal variational autoencoder (VAE) trained with neuroimaging and microRNA data. Our working hypothesis is that disease progression scores may be modelled as underlying latent traits. Concretely, we aim to learn a meaningful latent space, where the relative positions of latent representations indicate the distance travelled along the disease pathophysiological pathway.

VAEs are powerful generative models that project data into a low-dimensional regularized latent space [31]. These models have been previously used with multimodal data [32]–[34], but not for the purpose of inferring a DPS. Usually VAEs are trained in an unsupervised manner. However, extensions have been proposed for semi-supervised [35]–[37] or supervised [38] tasks. These studies demonstrate that providing supervision to the model imposes specific semantics on the latent space, resulting in more meaningful and robust representations. In our context, explicit labels (control, presymptomatic, patient) are already available for all subjects. We thus add supervision during training, leveraging this information to improve the separation of the groups in the latent space. Additionally, we propose to split high-dimensional (neuroimaging and microRNA data) and low-dimensional (demographic information) modalities. Our model thus couples two neural networks with different inputs: (1) an encoder/decoder that learns a latent space from the high-dimensional features, and (2) a classifier having as input the latent variables concatenated with the low dimensional features, useful for the classification task. As no ground truth is available for the DPS in real-world scenarios, we evaluate our models with a proxy metric: the area under the ROC curve (AUC) for each pairwise classification between clinical groups, computed using only the inferred DPS.

The main contributions of our paper are as follows. First, we propose a novel disease progression modeling framework that can be applied to cross-sectional data. Second, we introduce a supervised VAE to estimate a latent space that contains useful information for disease progression. To the best of our knowledge, supervised VAEs have never been used for disease progression modeling and have only been introduced in very different fields such as robotics [38]. Third, we propose to compute a disease trajectory in the latent space using principal curves. Fourth, we conduct very extensive experiments including: ablation studies to assess the importance of each new methodological component, different settings to study the robustness of the framework, and comparisons with state-of-the-art methods. A preliminary version has been published at the SPIE MI 2022 conference [39]. Compared to the conference version, the present paper introduces many novelties including VAE supervision, the use of principal curves instead of straight lines, and an extensive set of experiments.

The manuscript is organized as follows. Section II explains our proposed framework, section III describes the analyzed datasets, section IV details our experiments and corresponding outcomes, and finally section V examines the meaning of our results and highlights the broader implications of our study.

II. Methodology

We consider a dataset $(\mathcal{X}, \mathcal{Y}) = \{(x_1, y_1), ..., (x_n, y_n)\}$. The i-th subject is characterized by a feature vector $x_i \in \mathbb{R}^m$ and

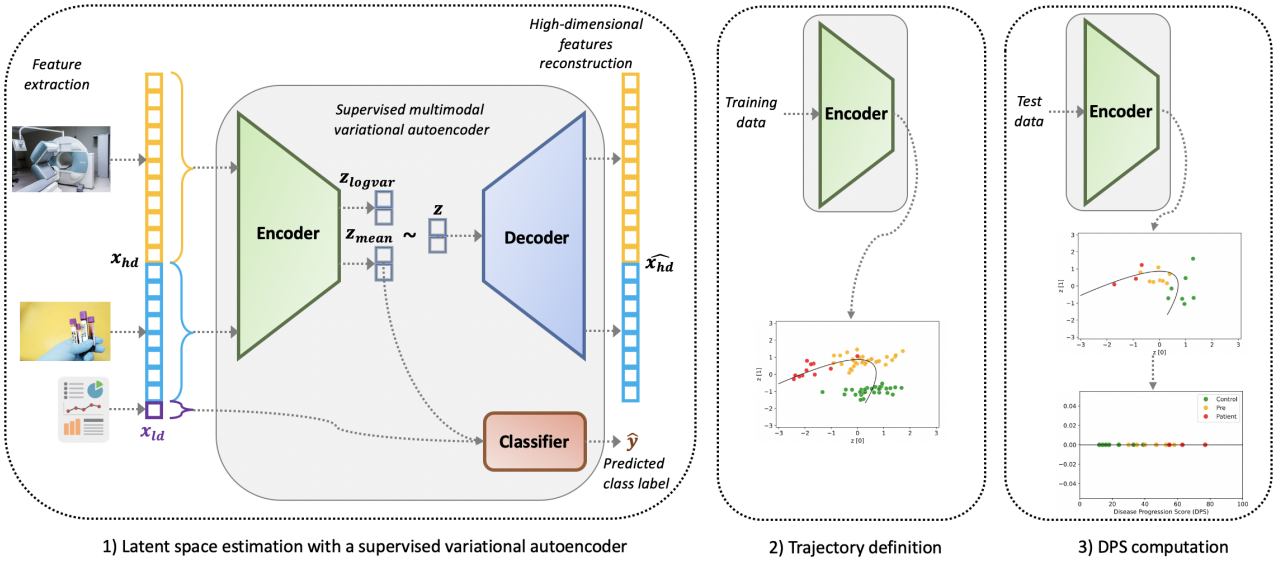


Fig. 1: Illustration of the proposed framework for disease progression scores (DPS) computation. 1) High-dimensional (neuroimaging and microRNAs expression data) and low-dimensional (demographic information) features are extracted; the former are fed to the encoder, the latter are concatenated with latent codes and fed to the classifier. 2) Once the model is trained, all training examples are encoded in the latent space and a principal curve is calculated to define the disease trajectory. 3) Test examples are encoded in the latent space and the latent representations are orthogonally projected onto the previously computed curve; the DPS correspond to their coordinates along the curve.

a label $y_i \in \{0, 1, 2\}$ denoting the clinical group (control, presymptomatic, patient). Our aim is to estimate a DPS, denoted as $v_i \in [1, 100]$ (the interval for the scores is arbitrary), where a greater score corresponds to a higher disease severity. To that purpose, we assume that the observations have corresponding latent variables $z_i \in \mathcal{R}^\ell$. We will thus aim to estimate a latent representation and the DPS will be computed from a trajectory in the latent space.

Our framework is composed of three main steps, as illustrated in Fig. 1. First, we propose a supervised multimodal variational autoencoder to estimate the latent space. We leverage the fact that participants belong to different groups to introduce some supervision in order to improve the VAE training. The model aims at simultaneously reconstructing the data and classifying the participants. We propose to split low-dimensional sociodemographic data (denoted $\mathcal{X}_{ld}$, used only for the classification) from high-dimensional multimodal neuroimaging and transcriptomic data (denoted $\mathcal{X}_{hd}$, used both for reconstruction and classification). Second, we build a curve representing disease trajectory in the latent space. Finally, data from new subjects, not included in the training set, are encoded in the latent space and projected onto this trajectory, in order to obtain their DPS.

In this section, we first explain the three main steps of our framework, then we describe implementation details.

A. Supervised multimodal VAE

A variational autoencoder (VAE) [31] is a generative model that learns the training data distribution $p(x)$ using a latent representation model: $p(x) = \int p(x|z)p(z)dz$, where z is a continuous latent variable living in a lower dimensional space and $p(z)$ is its prior distribution, commonly a Gaussian with zero mean and identity covariance matrix. The solution of the inference problem to describe the latent space is given by deriving the posterior $p(z|x)$. However, there is no closed-form solution for complex real-world datasets. Therefore, VAEs introduce the idea of learning a variational approximation $q_\phi(z|x)$ of the true posterior, in the form of a neural network referred to as the *encoder*. The encoder maps data x to a mean vector z_{mean} and a log-variance vector z_{logvar}, that parametrize a Gaussian distribution from which we obtain the latent representation z. VAEs are also equipped with a generative function $p_\theta(x|z)$, parametrized by a neural network referred to as the *decoder*. The decoder transforms the latent representation z back to the original input space.

During training, the vanilla VAE aims at maximizing the variational lower bound of the marginal log-likelihood, known as the evidence lower bound (ELBO). This is equivalent to minimizing a loss function with two terms: $\mathcal{L} = \mathcal{L}_r(x, \hat{x}) + \mathcal{L}_{KL}(q_\phi(z|x), p(z))$. The first term is the reconstruction error between the input data x and the reconstructed data $\hat{x}$, typically a mean squared error (MSE). The second term is the Kullback-Leibler divergence between the approximated posterior $q_\phi(z|x)$ and the prior distribution $p(z)$, acting as a regularization term.

We propose to insert a supervised branch in the vanilla VAE architecture in order to exploit the fact that our samples have different diagnostic labels, even though their DPS is unknown. Denoting y as the true class label and $\hat{y}$ as the predicted class label, we define our training objective as: $\mathcal{L} = \alpha_1 \cdot \mathcal{L}_r(x, \hat{x}) + \alpha_2 \cdot \mathcal{L}_{KL}(q_\phi(z|x), p(z)) + \alpha_3 \cdot \mathcal{L}_c(y, \hat{y})$, where $\mathcal{L}_r$ and $\mathcal{L}_{KL}$ correspond to the ELBO in vanilla VAEs and $\mathcal{L}_c$ is a cross-entropy term that penalizes the classification

error. The hyperparameters α_k control the relative weights between the different loss terms ($\sum_{k=1}^{3} \alpha_k = 1$).

Before training, we split the high-dimensional modalities (miRNA expression and neuroimaging) from the low-dimensional (demographic information). As it will be mentioned later in the datasets description, we consider one low-dimensional feature and $m-1$ high-dimensional features, although the same concepts can be applied to more low-dimensional features. So we use $m-1$ features to feed the encoder and one feature concatenated to the latent code to feed the classifier. Features are rescaled from 0 to 1. Our encoder consists of fully-connected layers of sizes $(m-1) \rightarrow 50 \rightarrow 2$, meaning our latent space is 2-dimensional. The decoder is implemented with fully-connected layers of sizes $2 \rightarrow 50 \rightarrow (m-1)$. The nonlinear activation function is the leaky rectified linear unit (ReLU) in all layers except the decoder's last layer which uses a sigmoid function to constrain the output between 0 and 1. The classifier network has one fully connected layer of $3 \rightarrow 3$ units, with a softmax function to normalize the output to probabilities over the predicted classes. We use the mean squared error as the reconstruction loss $\mathcal{L}_r$ and the cross-entropy as the classification loss $\mathcal{L}_c$.

This specific architecture, with a 2-dimensional latent space, has shown to be adequate for our real-world dataset. The proposed method is generic, and higher-dimensional latent spaces could yield better results with other datasets.

B. Trajectory definition

Once the model is trained, the next step is to encode the training data in the latent space. We then compute the straight line passing through the centroids of the control and patient clusters. This straight line could be used in downstream analyses as a rudimentary disease trajectory in the latent space. Instead, we obtain an improved nonlinear trajectory by using this line as initialization for the principal curve algorithm [40]. A principal curve is a smooth one-dimensional curve passing through the *middle* of given data points. The algorithm detailed in [40] finds a nonparametric curve by iteratively minimizing the orthogonal distances to the points until convergence.

C. DPS computation

Once the disease trajectory curve is computed in the latent space, we can encode the test data. The next step is to orthogonally project the latent codes onto the computed curve. The DPS $v_i \in [1, 100]$ for each subject is the coordinate of their projection along this curve, 1 corresponding to the beginning and 100 to the end of the curve. The pseudo-code from model training to DPS computation is shown in Algorithm 1.

D. Implementation details

The hyperparameters of the training objective were set as $\alpha_1 = 0.2$, $\alpha_2 = 0.2$, and $\alpha_3 = 0.6$. The loss function was optimized using Adam [41], with a learning rate of 10^{-3}, batches of 32 observations and 250 epochs.

We carried out the experiments on a computer equipped with a 2.4 GHz Intel Quad-Core Core i5 processor and 16 GB of RAM. Models were implemented in Python 3.8.5 using PyTorch 1.8.1 and Scikit-learn 0.23.2 [42]. For the principal curves computation, we used the implementation provided in the Python package pcurvepy 0.0.10 (`https://pypi.org/project/pcurvepy/`), specifying 2 as the degree of the smoothing spline.

Algorithm 1 DPS computation from latent representation

Input: features $\mathcal{X} = \{x_i\}_{i=1}^{n} \in \mathbb{R}^m$, labels $\mathcal{Y} = \{y_i\}_{i=1}^{n} \in \{0, 1, 2\}$, training set indices I_{tr} and test set indices I_{te} for one data split into training and test set.
Output: DPS $\{v_i\}_{i=I_{te}}$ of the subjects in the test set.
/* first step: supervised VAE training */
for epoch in [1,250] **do**
  Sample batches $(\mathcal{X}_j, \mathcal{Y}_j)$ from $(\mathcal{X}_{I_{tr}}, \mathcal{Y}_{I_{tr}})$
  for each batch $(\mathcal{X}_j, \mathcal{Y}_j)$ **do**
    $\mathcal{X}_{hd}, \mathcal{X}_{ld} \leftarrow$ split_high_low_dimension$(\mathcal{X}_j)$
    $\mathcal{Z}_{mean}, \mathcal{Z}_{logvar} \leftarrow$ encoder$(\mathcal{X}_{hd})$
    Draw latent codes $\mathcal{Z} \sim \mathcal{N}(\mathcal{Z}_{mean}, e^{\mathcal{Z}_{logvar}})$
    $\hat{\mathcal{Y}}_y \leftarrow$ classifier(concatenate$(\mathcal{X}_{ld}, \mathcal{Z}_{mean})$)
    $\hat{\mathcal{X}}_{hd} \leftarrow$ decoder$(\mathcal{Z})$
    $\mathcal{L}_r \leftarrow$ mean_squared_error$(\mathcal{X}_{hd}, \hat{\mathcal{X}}_{hd})$
    $\mathcal{L}_{KL} \leftarrow$ kl_divergence$(\mathcal{N}(\mathcal{Z}_{mean}, e^{\mathcal{Z}_{logvar}}), \mathcal{N}(0, I))$
    $\mathcal{L}_c \leftarrow$ cross_entropy$(\mathcal{Y}_y, \hat{\mathcal{Y}}_y)$
    $\mathcal{L} \leftarrow \alpha_1 \cdot \mathcal{L}_r + \alpha_2 \cdot \mathcal{L}_{KL} + \alpha_3 \cdot \mathcal{L}_c$
    Compute gradients, update network to minimize $\mathcal{L}$
  end for
end for
/* second step: trajectory definition */
$\mathcal{Z}, \_\_ \leftarrow encoder(\mathcal{X}_{I_{te}})$
$c_{control} \leftarrow$ mean$(\{\mathcal{Z}_j : y_j == 0\})$
$c_{patient} \leftarrow$ mean$(\{\mathcal{Z}_j : y_j == 2\})$
$pc \leftarrow$ principal_curve$(c_{control}, c_{patient}, degree = 2)$
/* third step: DPS computation */
for i in I_{te} **do**
  $z_{pc} \leftarrow$ projection of z_i into pc
  $v_i \leftarrow$ coordinate of $z_{pc} \in [0, 100]$
end for
return $\{v_i\}_{i=I_{te}}$

III. Datasets

A. Synthetic datasets

Since ground truth disease progression scores are not available in real-world scenarios, we created synthetic datasets to better evaluate the proposed framework. Multiple datasets were generated, with different noise levels and distinct proportions of features correlating with the DPS.

Let $Y \in \{0, 1, 2\}$ indicate the class labels (respectively control, presymptomatic and patient). We created $n = 111$ synthetic participants (a number close to that of our real dataset), denoting class labels by y_i $(i = 1, ..., 111)$, with $y_{i=1,...,37} = 0, y_{i=38,...,74} = 1, y_{i=75,...,111} = 2$.

Next, we modeled the disease progression scores as continuous random variables following uniform distributions. Let $V \in [1, 100)$ represent the DPS values. We defined the conditional distribution of the DPS given the class labels as

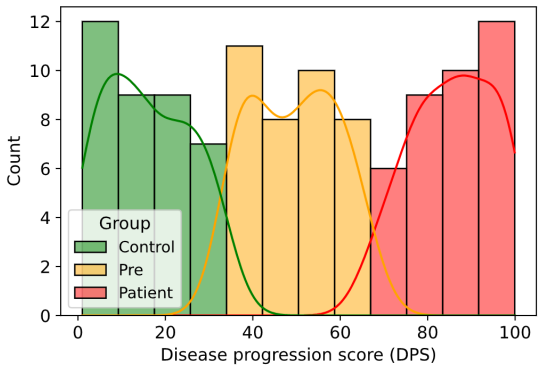


Fig. 2: Synthetic ground truth disease progression scores $\{v_i\}_{i=1}^{n} \in [0, 100)$ for $n = 111$ subjects (37 subjects per group).

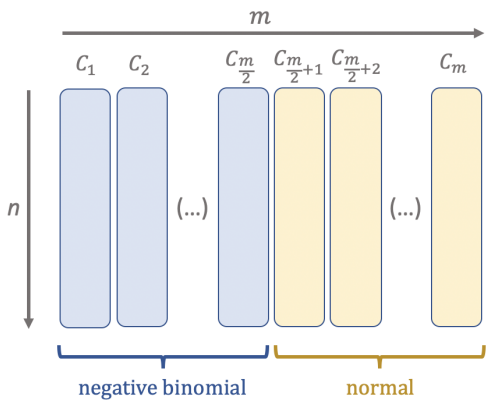


Fig. 3: Format of the synthetic datasets $\mathcal{D} \in \mathbb{R}^{n \times m}$ containing m features from n individuals. Half of the features are initially sampled from a negative binomial distribution and half from a normal distribution.

follows: $V|Y = 0 \sim \mathrm{U}[1, 34); V|Y = 1 \sim \mathrm{U}[34, 67); V|Y = 2 \sim \mathrm{U}[67, 100)$.

We then sampled the corresponding DPS v_i from the conditional distributions defined above. The obtained disease progression scores are displayed in Fig. 2.

Once the synthetic ground truth DPS were created, we generated multiple datasets $\mathcal{D} \in \mathbb{R}^{n \times m}$ containing $n = 111$ participants and $m = 160$ features. In order to simulate two modalities, features were initially sampled from two distributions: half from a negative binomial distribution (typical of miRNA expression data) and half from a normal distribution (representative of various real-world datasets). We denote the columns of $\mathcal{D}$ by $C_1, ..., C_m$. The format of the synthetic datasets is illustrated in Fig. 3.

Each created dataset had a distinct proportion of features correlating with the DPS and different noise levels. The number of features from each modality to positively and negatively correlate with the DPS is denoted as f, and the standard deviation of the added zero-mean Gaussian noise as s. We used $f = \{0, 2, 5, 10, 15, 20, 25, 30, 35, 40\}$ and $s = \{0.001, 0.2, 0.5, 0.8, 1, 5\}$ and thus obtained a total of 60 synthetic datasets.

B. Real dataset

Participants were recruited through the PREV-DEMALS study (https://clinicaltrials.gov, ID NCT02590276), a French multicentric prospective cohort focused on *C9orf72* expansion carriers. Written informed consents were obtained from all participants. The study was approved by the ethics committee (Comité de Protection des Personnes CPP Ile-De-France VI, CPP 68-15 and ID RCB 2015-A00856-43). A detailed description of this cohort and its demographic profile can be found in [5].

We included 110 individuals in our analyses, divided into three groups, according to their clinical status:

- Patient group: 22 symptomatic (15 FTD, 4 FTD/ALS and 3 ALS) carriers of a pathogenic *C9orf72* expansion;
- Presymptomatic group: 45 asymptomatic carriers;
- Control group: 43 asymptomatic non-carriers.

The dataset comprised multimodal data including microRNA (miRNA) sequencing data and neuroimaging data. Among the 110 subjects, 91 had complete miRNA expression and neuroimaging data, while 19 had only miRNA expression available. The two modalities are described below.

1) MicroRNA data: MicroRNAs are a class of small non-coding RNAs that negatively regulate gene expression [43]. MicroRNAs expression in blood plasma has been shown to correlate with the diagnosis and progression of many neurodegenerative diseases [44], including FTD and ALS. All individuals included in this cohort underwent plasma sampling, from which miRNA sequencing was performed. Plasma collection and preparation, miRNA extraction and sequencing, quality control and the computational pipeline to obtain the miRNA counts are detailed in [5]. The initial miRNA dataset contained expression levels for all miRNAs mapped in the human genome (2576 miRNAs). We retained the 589 miRNAs with expression profiles above noise level (minimum total count of 1000 reads and at least 50 reads for one sample). A trimmed mean of M-values [45] implemented in the R package EdgeR [46] was used to normalize the raw counts.

2) Neuroimaging data: Neuroimaging data consisted of grey matter volumes extracted from T1-weighted anatomical magnetic resonance imaging (MRI), including the estimated total intracranial volume (TIV), 68 cortical regions of interest (ROIs) using the Desikan atlas and 18 subcortical ROIs using the Aseg nomenclature, thus resulting in 87 neuroimaging features. The TIV was used to normalize the volume of each ROI,

$$NV_{ROI} = \frac{TIV_m \times V_{ROI}}{TIV},$$

where V_{ROI} is the original volume of the ROI, NV_{ROI} is the corresponding normalized volume and TIV_m is the average TIV computed across all subjects. The MRI acquisition parameters, quality check and processing pipeline are thoroughly described in [4].

Since only 91 subjects (14 patients, 40 presymptomatic carriers and 37 controls) had MRI scans collected, we divided our dataset into two subsets: 19 subjects that only had miRNA data available, and 91 subjects with multimodal neuroimaging and miRNA data. The former subset was used as a discovery

set for miRNA feature selection: we used these 19 individuals to perform differential expression analysis (as described in [5]). The 68 miRNAs with the lowest p-values were selected for all downstream analyses.

Lastly, we also included age as demographic information for all subjects. So the total dimension of each feature vector was $m = 87 + 68 + 1 = 156$.

IV. Experiments and Results

A. Synthetic datasets

We applied our framework to 60 synthetic datasets (described in Section III-A) with different noise levels and distinct number of features correlating with the ground truth DPS. Each synthetic dataset was divided into a training set of 90 subjects (30 per clinical group) and a test set of 21 individuals (7 per group). We trained one model per dataset, using the same hyperparameters as the experiments with the real dataset. After training each model, we computed the DPS for the subjects from the test set. We then calculated the Spearman correlations between the simulated ground-truth scores and the predicted scores. Finally, we evaluated the ROC AUC for each pairwise comparison between the three simulated clinical groups.

Fig. 4 presents the computed trajectories and the DPS obtained when 50% of the features are correlated (25% positively and 25% negatively correlated) with the disease progression, for six different noise levels. The correlation matrices illustrate the strength of the relationships between the simulated features, for all investigated noise levels. Lower noise levels (Fig. 4a-c) imply a strong correlation between the simulated features and DPS, which in turn result in more meaningful inferred trajectories and finally more accurate estimated DPS.

The results of the Spearman correlation between the synthetic ground truth and the estimated DPS, as well as the average ROC AUC scores for the three pairwise comparison between groups, are shown in Fig. 5, for six different noise levels. We note that two factors imply higher Spearman correlations, and thus more accurate DPS estimations: first, lower noise levels (Fig. 5 top rows); and second, a higher proportion of relevant features (Fig. 5 horizontal axes). Importantly, we observe that the Spearman correlation of the DPS and the macro-average ROC AUC have similar behaviors, indicating that the ROC AUC of pairwise comparisons is a reasonable proxy to evaluate the DPS, as will be done with the real dataset.

Additionally, we conducted additional experiments with imbalanced datasets (fixing $f = 20$ and $s = 0.5$, following the notation from Section III). We computed the macro-average ROC AUC and the Spearman correlation between the ground truth and the estimated DPS, for two scenarios with moderate imbalance and two scenarios with strong imbalance. Results are displayed in Appendix I Table A.1. The model has a good performance with moderate imbalance (minor class with 10% of the size of the major class), but heavily underperforms with strong imbalance (minor class with 1% of the size of the major class).

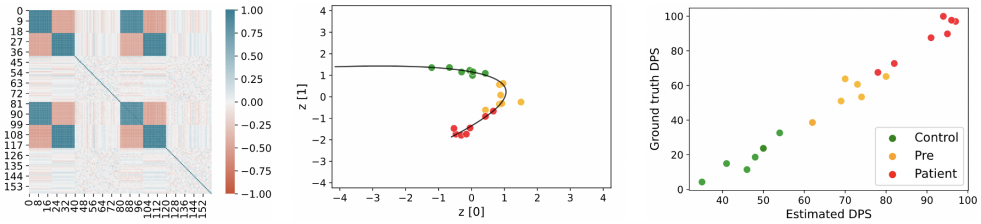
Finally, we analyzed the performance of our approach with higher dimensional synthetic datasets, containing 16000 features. In each experiment, we varied the number of features that correlated with the simulated ground truth DPS, while keeping the added noise fixed ($s = 0.5$). The macro-average ROC AUC and Spearman correlation between ground truth and estimated DPS are shown in Appendix I Table A.2. We note that the greater the number of features correlating with the DPS, the better the results. Therefore, the most important factor affecting performance is not the absolute number of features, but how many of these features are actually useful to estimate the DPS.

B. Real dataset

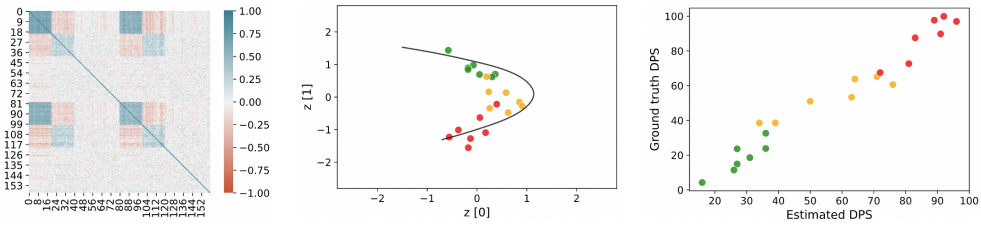
Experiments with the real dataset (described in Section III-B) were carried out with a cross-validation of 100 stratified randomized folds. For each fold split, we trained a model using 73 training subjects, and then computed the DPS for the 18 individuals in the test set. Fig. 6 displays an example of the DPS computation for one representative training data split. For each split, once the model was trained, the training data was encoded in the latent space and the disease trajectory was computed (Fig. 6a). Then, the corresponding test set was encoded in the latent space, and each latent code was projected onto the previously computed trajectory (Fig. 6b). Finally, disease progression scores were obtained for each subject as the coordinates along the trajectory (Fig. 6c).

Unlike for the synthetic dataset, there is no ground truth for the DPS in the real dataset. We thus applied a proxy metric to assess model performance: using only the inferred DPS, we did pairwise comparisons between the clinical groups and computed the corresponding areas under the ROC curves. Specifically, we present the following experiments: (1) evaluation of the proposed method, (2) comparison with the state-of-the-art methods for modeling disease progression with cross-sectional data (EBM), (3) ablation study, (4) variation of hyperparameters, (5) contribution of different modalities, and (6) impact of incorrect diagnosis.

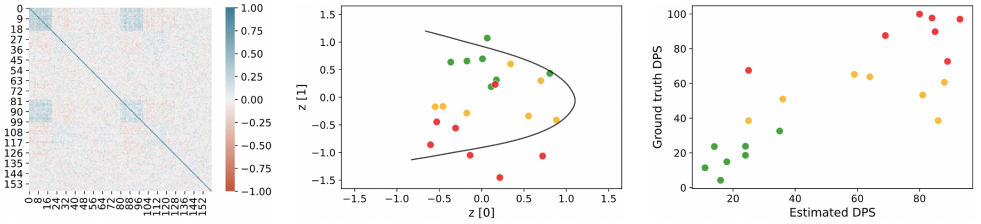
1) Evaluation of the proposed method: First, we used the DPS computed in each fold to build ROC curves for the three pairwise comparisons between clinical groups. The average ROC curves are shown in Fig. 7. The ROC AUC for the classification of controls and presymptomatic subjects was 0.74 ± 0.13, for controls and patients was 0.98 ± 0.05 and to distinguish presymptomatic carriers and patients was 0.96 ± 0.07. These results reveal that it is harder to differentiate controls from presymptomatic individuals than it is to distinguish between patients and the other two groups. The histogram displayed in Fig. 8 illustrates the disease progression scores computed over all 100 test folds (18 subjects per test fold, corresponding to 1800 DPS). The distribution shapes highlight a clear separation between the patient group and the other groups. The distribution of the DPS for the presymptomatic group is more spread, which was expected as this group is the most heterogeneous. Some presymptomatic subjects are very far from onset and the neurodegenerative process has barely begun, they are thus closer to controls. Other presymptomatic subjects are closer to disease onset and thus their DPS is closer to that of patients. Finally, the Spearman correlation computed



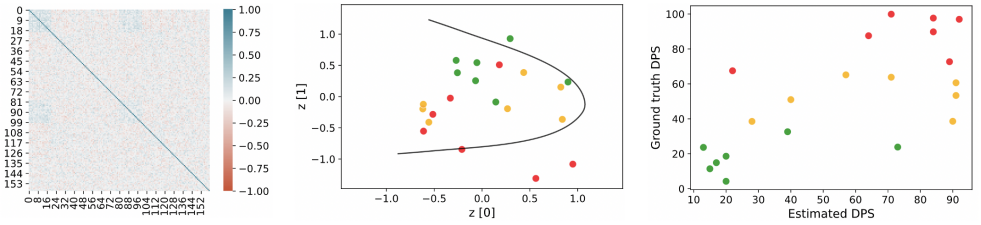
(a) Noise standard deviation = 0.001



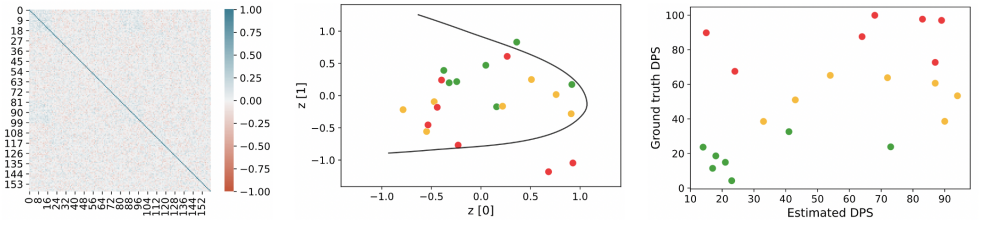
(b) Noise standard deviation = 0.2



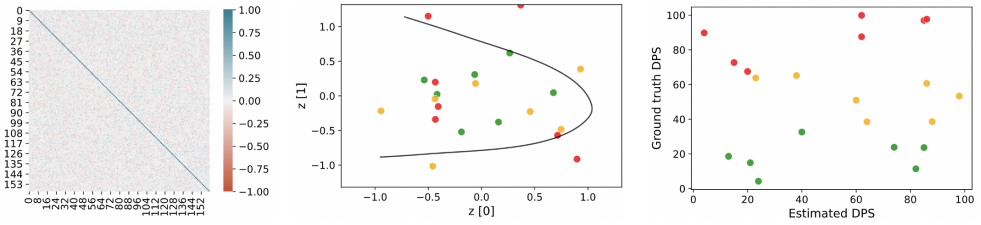
(c) Noise standard deviation = 0.5



(d) Noise standard deviation = 0.8



(e) Noise standard deviation = 1



(f) Noise standard deviation = 5

Fig. 4: Results on synthetic data when 50% of the features are correlated with the disease progression score. The rows indicate different noise levels (zero-mean Gaussian noise with different standard deviations). Each column displays, respectively: (1) correlation matrices showing the strength of the relationships between the simulated features, (2) inferred trajectories and test sets projected in the latent space, and (3) estimated DPS vs. ground truth DPS.

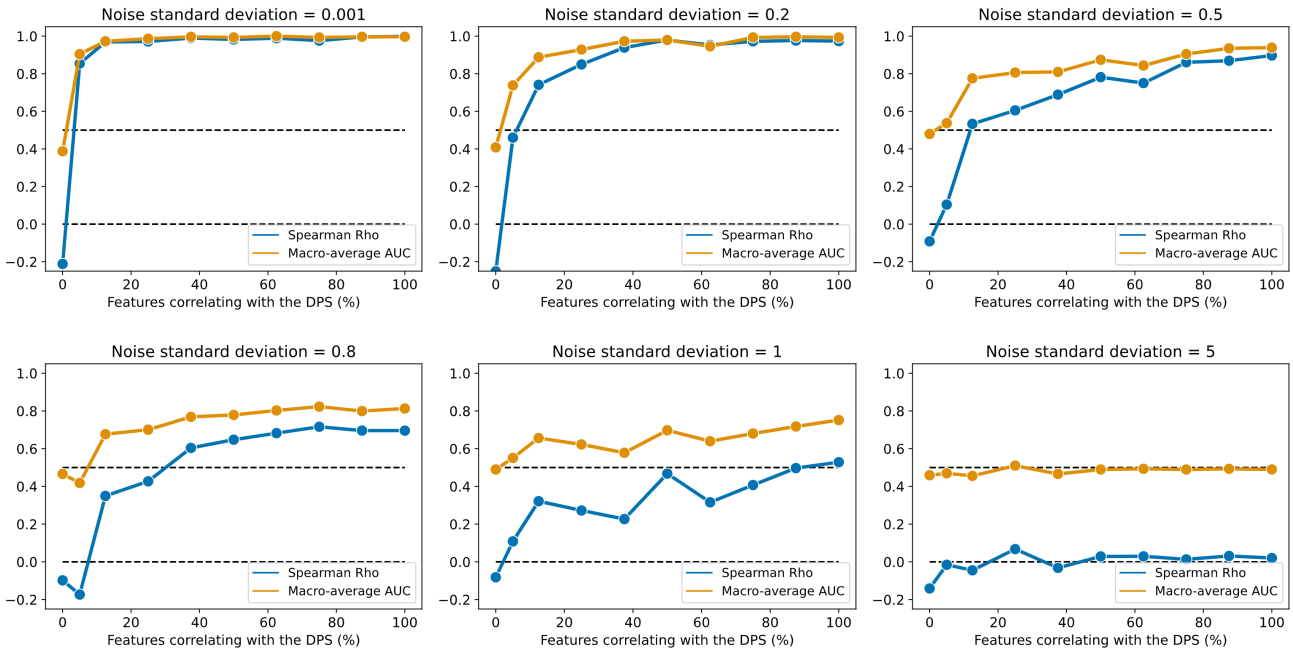
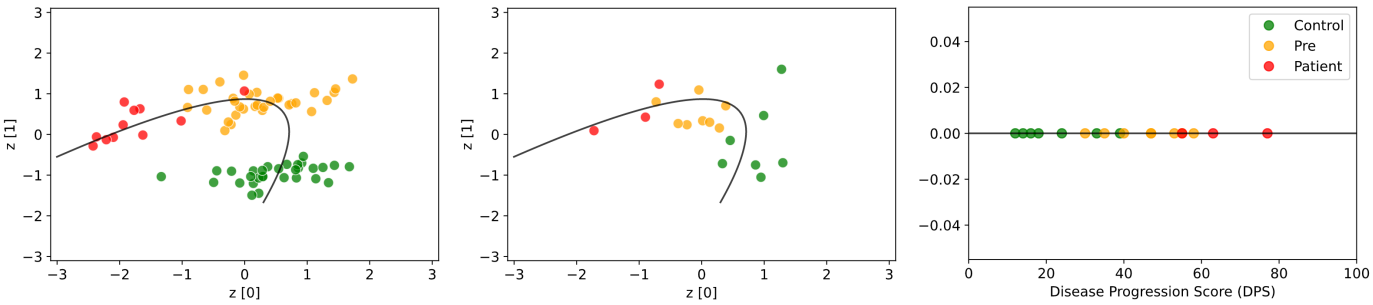


Fig. 5: Results on synthetic data. Macro-average ROC AUC and Spearman correlation between ground truth and estimated DPS, for different noise levels (zero-mean Gaussian with 0.001, 0.2, 0.5, 0.8, 1, and 5 as standard deviation) and several proportions (0% to 100%) of features correlating with the disease progression score. Random chance is denoted by the dashed lines (ROC AUC = 0.5 and Spearman Rho = 0).



(a) Training data projected in the latent space (b) Test data projected in the latent space (c) Computed scores for the test data

Fig. 6: Results on real data. (a) Training data projected in the latent space and the corresponding computed trajectory for one of the 100 fold splits. (b) Test data projected in the latent space, along with the previously computed trajectory. (c) Scores computed after the projection of the latent representation of the test data onto the trajectory.

between DPS and age was 0.54 for the presymptomatic group and 0.38 for the patient group. It is expected that DPS and age are somewhat correlated, since older subjects tend to have more advanced neurodegeneration.

2) Comparison with EBM: Next, we compared our results to three different event-based models, which are the state-of-the-art methods to model disease progression from cross-sectional data: two discriminative event-based models (DEBM) [8], [9] and one generative EBM [7] that extended the original version of the algorithm [6]. For that experiment, the same cross-validation strategy of 100 stratified folds was applied. We built all the event-based models and computed the DPS using the Python package pyebm 2.0.3 (`https://pypi.org/project/pyebm/`), optimizing for the best staging algorithm among the four choices offered by the package. Table I displays the corresponding ROC AUC results for each pairwise comparison, while the ROC curves obtained with our proposed approach are displayed alongside the ROC curves yielded by the best performing EBM in Fig. 7. We can observe that our model achieves a substantially better classification performance for all pairwise comparisons. Additionally, our approach used less computing time: our framework took 2 seconds per fold for training and DPS computation, while the EBM algorithms took on average 180 seconds per fold.

3) Ablation study: Afterwards, to investigate the impact of certain components of our framework, we conducted an ablation study. We changed some elements of the proposed approach to obtain four alternative models:

- Linear trajectory: rather than computing the trajectory in the latent space using principal curves, we simply used a straight line.
- No supervision: we removed the entire classification

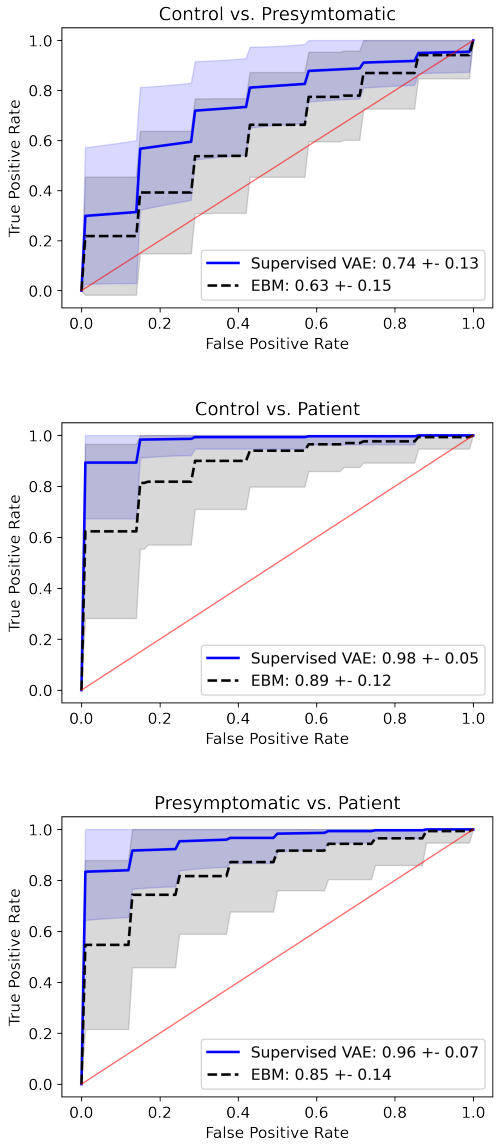


Fig. 7: Results on real data. Average ROC (receiver operating characteristic) curves for each pairwise comparison between clinical groups, over 100 stratified splits, for our proposed supervised VAE and for the best performing EBM [7]. The shaded areas correspond to one standard deviation. The areas under the ROC curves (ROC AUC) are shown as mean ± standard deviation. Random chance is indicated by the red line.

component of our framework, thus performing unsupervised training, and discarding the low-dimensional demographic information.

- Concatenated modalities: we concatenated the low-dimensional demographic information with the high-dimensional modalities in the encoder input, and used only the latent codes as input for the classifier.
- No supervision, concatenated modalities: we removed the entire classification branch and concatenated the low and high-dimensional modalities.

For each alternative model, we conducted the same cross-validation strategy of 100 stratified folds, computing the DPS for the test sets and the corresponding areas under the ROC curves. The results, displayed in Table II, show that the proposed model has a better and more stable performance in all comparisons, with the highest average ROC AUC and lowest standard deviation among the splits.

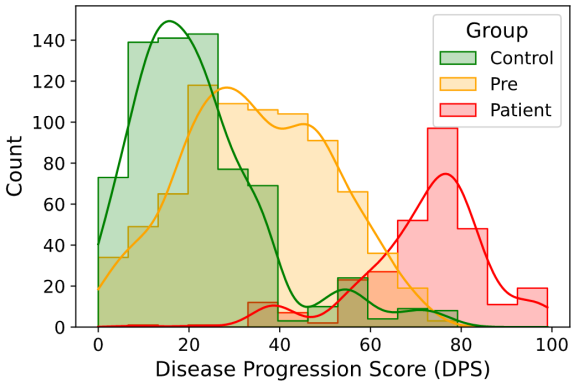


Fig. 8: Results on real data. Histogram of the disease progression scores (DPS) inferred for 18 test subjects over 100 stratified splits. The distribution shapes are approximated with kernel density estimates.

TABLE I: Results on real data: comparison between our approach, two discriminative event-based models (DEBM) [8], [9], and a generative event-based model (EBM) [7]. ROC AUC (mean ± standard deviation) over 100 stratified splits.

Comparison	Our model	DEBM [8]	DEBM [9]	EBM [7]
Control vs. Pre	**0.74 ± 0.13**	0.65 ± 0.16	0.68 ± 0.14	0.63 ± 0.15
Control vs. Patient	**0.98 ± 0.05**	0.80 ± 0.14	0.78 ± 0.17	0.89 ± 0.12
Pre vs. Patient	**0.96 ± 0.07**	0.71 ± 0.17	0.68 ± 0.17	0.85 ± 0.14

4) Variation of hyperparameters: We checked whether our results were robust to reasonable changes in the hyperparameters. Notably, we tested different numbers of hidden units in the fully-connected layers, and different combinations of the relative weights between the loss terms. These results are summarized in Appendix I Tables A.3 and A.4. The slightly different but overall similar results demonstrate that our hyperparameter choice is not overfitting the data.

5) Contribution of different modalities: Appendix I Table A.5 shows the ROC AUC results using each modality separately. Notably, microRNAs are more informative than neuroimaging to distinguish between controls and presymptomatic individuals, while neuroimaging is more informative to classify between patients and the two other groups. Additionally, age alone achieves a high ROC AUC when classifying between patients and the two other groups, which is expected, as the patients are older than the other participants.

6) Impact of incorrect diagnosis: Finally, to assess the impact of incorrect diagnosis, we performed 100 stratified splits randomly changing 10% of the training labels. Results are displayed in Appendix I Table A.6. Although the classification branch depends on diagnoses as labels, this is just one component of the model, and a few label errors did not considerably impact results.

TABLE II: Results on real data: ablation study. ROC AUC results (mean ± standard deviation) for the proposed model and four alternative models from the ablation study.

Comparison	Proposed model	Linear trajectory	No supervision	Concatenated modalities	No supervision, concatenated modalities
Control vs. Pre	**0.74 ± 0.13**	0.62 ± 0.15	0.67 ± 0.15	0.72 ± 0.15	0.65 ± 0.15
Control vs. Patient	**0.98 ± 0.05**	0.93 ± 0.12	0.96 ± 0.06	0.95 ± 0.17	0.96 ± 0.06
Pre vs. Patient	**0.96 ± 0.07**	0.93 ± 0.11	0.94 ± 0.10	0.91 ± 0.18	0.94 ± 0.10

V. Discussion

In this paper, we proposed a new approach for estimating disease progression scores from cross-sectional neuroimaging and transcriptomic data that is applicable in small samples, which are typically found in rare diseases. The approach was designed and evaluated on data from *C9orf72*-associated FTD and ALS, but is potentially applicable to other diseases. Results on synthetic data demonstrated the ability of the method to accurately estimate the DPS, and experiments on real data, in the absence of ground truth DPS, showed the separation of different diagnostic classes. The findings of this study validated the usefulness of supervised variational autoencoders to infer disease trajectories from cross-sectional multimodal data, indicating that a single disease progression score may be used to represent progression of neurodegenerative diseases. Remarkably, our results revealed that the DPS may be inferred using only cross-sectional data from a small sample of subjects.

Experiments with a cohort of *C9orf72*-mutation carriers demonstrate that subjects from the same clinical groups (patients, presymptomatic individuals and controls) are clustered together in the latent space (Fig. 6), allowing the inference of a disease trajectory. After training the model, data from new individuals is encoded in the latent space and orthogonally projected onto this trajectory to compute the DPS. Notably, using only the computed DPS, we are able to classify presymptomatic subjects and patients with an average ROC AUC of 0.96 over 100 stratified fold splits (Fig. 7), illustrating how much the DPS reflects the degree of disease progression in mutation carriers. Unsurprisingly, it is harder to differentiate between controls and presymptomatic individuals, as indicated by the average ROC AUC of 0.74 and displayed in Fig. 8. This stems from the fact that, during earlier disease stages, most biomarker levels are closer to normal ranges, so the presymptomatic class is more heterogeneous.

To the best our of knowledge, event-based models are the only published methods to compute disease progression scores from cross-sectional data, other approaches requiring longitudinal data. The comparisons presented in Table I reveal that our proposed approach resulted in considerably higher ROC AUC than EBM for all pairwise classifications. This suggests that the supervised VAE is more suitable than event-based models for DPS computation with high-dimensional features, such as multimodal microRNA and neuroimaging data. Indeed, published studies using event-based models explored a substantially lower number of features. For instance, in Alzheimer's disease, EBM experiments were carried our with 13 to 50 [6]–[9], [21]–[23] biomarkers. Studies focusing on FTD analyzed 21 [28] or 7 [29] biomarkers, while multiple sclerosis was investigated with 25 [24] or 24 [25] biomarkers. Other conditions such as Parkinson's disease [26], ALS [30] and Huntington's disease [27] were modeled with respectively 42, 19 and 8 biomarkers. Nevertheless, the EBM model presents useful additional features, beyond the computation of DPS. In particular, it can provide a temporal ordering of when the different biomarkers become abnormal, which is useful for understanding disease progression. Moreover, a balance has to be found between the number of features and the number of subjects in each dataset. Indeed, we also had to perform feature selection to decrease the number of microRNAs in our study. It should be noted that this feature selection was unbiased, since it was performed using a completely separate set of participants that was not used in the rest of the study. The proposed framework was able to achieve a good performance with 156 features and less than a hundred subjects, thus demonstrating its potential for dealing with higher dimensional datasets.

An ablation study evaluated the impact of different components of our approach (Table II). We observed that each component positively impacted the framework's performance. First, it can be seen that a curved trajectory better fits the disease pathway in the latent space when compared to a straight line. The use of principal curves has been inspired from their application in a similar task: pseudotime inference for single-cell transcriptomics, as shown in [47]. In that context, pseudotime represents an underlying temporal variable driving a smooth transition between cellular states, and principal curves are used to infer a trajectory in a low-dimensional space. Second, it is clear that the addition of supervision with a classifier branch improves the separation between clinical groups in the latent space. Rather than discrete clusters, our experiments demonstrate that latent representations are placed along a continuous path. Specifically, supervision adds meaning to the relative positions between points in the latent space. Finally, results show the contribution of splitting high and low-dimensional features. When using the low-dimensional features concatenated with the latent codes as inputs to the classifier, the model's performance is enhanced. The same pattern is observed in [38], although in a totally different context (failure detection in robotics). Concretely, a low-dimensional feature can directly contribute to the classifier, without the need for encoding.

Regarding the experiments with simulated datasets, it is crucial to highlight the relationship of the average ROC AUC with the Spearman correlation between ground truth and estimated DPS (Fig. 5). The simulation supports that the higher the ROC AUC, the more accurate the predicted DPS. Therefore, for real-world scenarios without ground truth DPS, our choice of the ROC AUC as proxy metric is corroborated.

Furthermore, evidence was found that the models do not overfit the data, since it is clear that larger noise levels lead to poorer results, eventually equivalent to random chance. The effect of noise is further illustrated in Fig. 4. We observe that lower noise levels induce more evident clusters and more meaningful trajectories in the latent space. Consequently, the estimated DPS are closer to the ground truth. These simulations also confirm one intuition behind our model: the more features correlate with disease progression, the closer the estimated DPS are to the ground truth.

Our study has the following limitations. First, there is no ground truth for the progression scores in real datasets. Although the experiments with synthetic data showed that the ROC AUC is an adequate proxy metric, long-term follow-up of individuals will be necessary to assess the accuracy of the computed DPS. For instance, we need follow-up data to confirm the hypothesis that a higher DPS implies an earlier disease onset for a presymptomatic subject. Second, the lack of a replication cohort means that additional studies will be necessary to further support the clinical relevance of our findings. Third, as it is the case with other supervised algorithms, our proposed method will underperform with strongly imbalanced datasets. Fourth, the interpretability of our approach is limited in comparison with event-based models, which directly estimate the ordering in which biomarkers become abnormal. Finally, our model does not deal with incomplete datasets: to infer a disease progression score for a new subject, all the features that were used during training are needed. To address incomplete datasets, possible solutions could be implementing multiple stages in the network, to gradually integrate available multimodal data in each stage [48], or even synthesizing data from the available modalities [49]. Future work will also concentrate on the integration of more data sources, such as positron emission tomography (PET) scans and neurofilament light chain (NfL) levels in blood.

In conclusion, we proposed a new approach to measure disease progression from multimodal imaging and microRNA data in rare neurodegenerative disorders using only cross-sectional data. Even though we focused on *C9orf72*-associated FTD and ALS, our framework is generic. It has the potential to be useful for a variety of other diseases, enabling the evaluation of novel treatments even when only cross-sectional data from small cohorts are available.

ACKNOWLEDGMENT

The PREV-DEMALS study group includes: Eve Benchetrit[1], Anne Bertrand[1], Anne Bissery[1], Marie-Paule Boncoeur[2], Stéphanie Bombois[3], Agnès Camuzat[4], Mathieu Chastan[5], Yaohua Chen[3], Marie Chupin[4], Olivier Colliot[4], Philippe Couratier[2], Xavier Delbeuck[3], Vincent Deramecourt[3], Christine Delmaire[3], Emmanuel Gerardin[5], Claude Hossein-Foucher[3], Bruno Dubois[1], Marie-Odile Habert[1], Didier Hannequin[5], Géraldine Lautrete[2], Thibaud Lebouvier[3], Isabelle Le Ber[1], Benjamin Le Toullec[4], Richard Levy[1], Olivier Martinaud[5], Kelly Martineau[4], Marie-Anne Mackowiak[3], Jacques Monteil[2], Florence Pasquier[3], Gregory Petyt[3], Pierre-François Pradat[1], Assi-Hervé Oya[1], Armelle Rametti-Lacroux[1], Daisy Rinaldi[1], Adeline Rollin-Sillaire[3], François Salachas[1], Sabrina Sayah[1], David Wallon[5].

[1]Hôpital de la Salpêtrière ; [2]CHU Dupuytren ; [3]CHU Roger Salengro ; [4]Paris Brain Institute ; [5]CHU Charles Nicolle .

APPENDIX I
SUPPLEMENTARY EXPERIMENTS

TABLE A.1: Macro-average ROC AUC and Spearman correlation between the ground truth and the estimated DPS, for simulated datasets with different class imbalance levels. Original results are shown in bold.

Subjects in each clinical group	**30, 30, 30**	300, 30, 30	300, 30, 300	3000, 30, 30	3000, 30, 3000
Macro-average AUC	**0.83**	0.81	0.82	0.74	0.67
Spearman Rho	**0.76**	0.70	0.65	0.50	0.30

TABLE A.2: Macro-average ROC AUC and Spearman correlation between the ground truth and the estimated DPS, for high-dimensional simulated datasets with different proportions of informative features. Original results are shown in bold.

Features (correlated)	**160 (80)**	16000 (80)	16000 (320)	16000 (640)	16000 (1280)
Macro-average AUC	**0.83**	0.58	0.82	0.85	0.99
Spearman Rho	**0.76**	0.32	0.65	0.72	0.99

TABLE A.3: ROC AUC results on real data (mean ± SD) over 100 stratified splits when changing the number of units of the hidden layers. Original results are shown in bold.

Hidden units	**50**	100	80	25
Control vs. Pre	**0.74 ± 0.13**	0.73 ± 0.13	0.71 ± 0.12	0.71 ± 0.13
Control vs. Patient	**0.98 ± 0.05**	0.98 ± 0.04	0.97 ± 0.05	0.98 ± 0.05
Pre vs. Patient	**0.96 ± 0.07**	0.96 ± 0.06	0.96 ± 0.06	0.96 ± 0.06

TABLE A.4: ROC AUC results on real data (mean ± SD) over 100 stratified splits when changing the weights of the loss function terms. Original results are shown in bold.

Weights α_k	**0.2, 0.2, 0.6**	0.1, 0.1, 0.8	0.1, 0.2, 0.7	0.3, 0.2, 0.5
Control vs. Pre	**0.74 ± 0.13**	0.72 ± 0.12	0.73 ± 0.12	0.72 ± 0.14
Control vs. Patient	**0.98 ± 0.05**	0.97 ± 0.08	0.97 ± 0.06	0.98 ± 0.05
Pre vs. Patient	**0.96 ± 0.07**	0.94 ± 0.10	0.95 ± 0.09	0.96 ± 0.07

TABLE A.5: ROC AUC results on real data (mean ± SD) over 100 stratified splits, analyzing the impact of each modality. The best results in each comparison are shown in bold.

Comparison	MicroRNA	Neuroimaging	Demographic
Control vs. Pre	**0.73 ± 0.12**	0.60 ± 0.15	0.40 ± 0.13
Control vs. Patient	0.81 ± 0.18	**0.97 ± 0.04**	0.81 ± 0.14
Pre vs. Patient	0.67 ± 0.19	**0.96 ± 0.05**	0.91 ± 0.09

TABLE A.6: ROC AUC results on real data (mean ± SD) over 100 stratified splits when changing 10% of the training labels. Original results are shown in bold.

Comparison	**Original results**	Changing 10% of training labels
Control vs. Pre	**0.74 ± 0.13**	0.72 ± 0.15
Control vs. Patient	**0.98 ± 0.05**	0.98 ± 0.06
Pre vs. Patient	**0.96 ± 0.07**	0.95 ± 0.08

REFERENCES

[1] M. DeJesus Hernandez *et al.*, "Expanded GGGGCC hexanucleotide repeat in noncoding region of C9ORF72 causes chromosome 9p-linked FTD and ALS," *Neuron*, vol. 72, pp. 245–256, Oct. 2011.

[2] A. E. Renton *et al.*, "A hexanucleotide repeat expansion in C9ORF72 is the cause of chromosome 9p21-linked ALS-FTD," *Neuron*, vol. 72, pp. 257–268, Oct. 2011.

[3] J. D. Rohrer *et al.*, "Presymptomatic cognitive and neuroanatomical changes in genetic frontotemporal dementia in the Genetic Frontotemporal dementia Initiative (GENFI) study: a cross-sectional analysis," *The Lancet. Neurology*, vol. 14, pp. 253–262, Mar. 2015.

[4] A. Bertrand *et al.*, "Early Cognitive, Structural, and Microstructural Changes in Presymptomatic C9orf72 Carriers Younger Than 40 Years," *JAMA neurology*, vol. 75, no. 2, pp. 236–245, 2018.

[5] V. Kmetzsch *et al.*, "Plasma microRNA signature in presymptomatic and symptomatic subjects with C9orf72-associated frontotemporal dementia and amyotrophic lateral sclerosis," *Journal of Neurology, Neurosurgery & Psychiatry*, vol. 92, pp. 485–493, May 2021.

[6] H. M. Fonteijn *et al.*, "An event-based model for disease progression and its application in familial Alzheimer's disease and Huntington's disease," *NeuroImage*, vol. 60, pp. 1880–1889, Apr. 2012.

[7] A. L. Young *et al.*, "A data-driven model of biomarker changes in sporadic Alzheimer's disease," *Brain*, vol. 137, pp. 2564–2577, Sept. 2014.

[8] V. Venkatraghavan *et al.*, "A Discriminative Event Based Model for Alzheimer's Disease Progression Modeling," in *Information Processing in Medical Imaging*, Lecture Notes in Computer Science, (Cham), pp. 121–133, 2017.

[9] V. Venkatraghavan *et al.*, "Disease progression timeline estimation for Alzheimer's disease using discriminative event based modeling," *NeuroImage*, vol. 186, pp. 518–532, Feb. 2019.

[10] B. M. Jedynak *et al.*, "A computational neurodegenerative disease progression score: Method and results with the Alzheimer's disease neuroimaging initiative cohort," *NeuroImage*, vol. 63, pp. 1478–1486, Nov. 2012.

[11] M. Mehdipour Ghazi *et al.*, "Robust parametric modeling of Alzheimer's disease progression," *NeuroImage*, vol. 225, p. 117460, Jan. 2021.

[12] J.-B. Schiratti *et al.*, "A Bayesian Mixed-Effects Model to Learn Trajectories of Changes from Repeated Manifold-Valued Observations," *Journal of Machine Learning Research*, vol. 18, no. 133, pp. 1–33, 2017.

[13] I. Koval *et al.*, "AD Course Map charts Alzheimer's disease progression," *Scientific Reports*, vol. 11, p. 8020, Apr. 2021.

[14] R. V. Marinescu *et al.*, "DIVE: A spatiotemporal progression model of brain pathology in neurodegenerative disorders," *NeuroImage*, vol. 192, pp. 166–177, May 2019.

[15] M. Lorenzi *et al.*, "Probabilistic disease progression modeling to characterize diagnostic uncertainty: Application to staging and prediction in Alzheimer's disease," *NeuroImage*, vol. 190, pp. 56–68, Apr. 2019.

[16] P. A. Wijeratne *et al.*, "Revealing the Timeline of Structural MRI Changes in Premanifest to Manifest Huntington Disease," *Neurology Genetics*, vol. 7, Oct. 2021.

[17] S. Garbarino and M. Lorenzi, "Investigating hypotheses of neurodegeneration by learning dynamical systems of protein propagation in the brain," *NeuroImage*, vol. 235, p. 117980, July 2021.

[18] S. Garbarino *et al.*, "Differences in topological progression profile among neurodegenerative diseases from imaging data," *eLife*, vol. 8, p. e49298, 2019.

[19] L. M. Aksman *et al.*, "Modeling longitudinal imaging biomarkers with parametric Bayesian multi-task learning," *Human Brain Mapping*, vol. 40, pp. 3982–4000, Sept. 2019.

[20] M. Mehdipour Ghazi *et al.*, "Training recurrent neural networks robust to incomplete data: Application to Alzheimer's disease progression modeling," *Medical Image Analysis*, vol. 53, pp. 39–46, Apr. 2019.

[21] N. P. Oxtoby *et al.*, "Data-driven models of dominantly-inherited Alzheimer's disease progression," *Brain*, vol. 141, pp. 1529–1544, May 2018.

[22] N. C. Firth *et al.*, "Sequences of cognitive decline in typical Alzheimer's disease and posterior cortical atrophy estimated using a novel event-based model of disease progression," *Alzheimer's & Dementia*, vol. 16, pp. 965–973, July 2020.

[23] D. Archetti *et al.*, "Multi-study validation of data-driven disease progression models to characterize evolution of biomarkers in Alzheimer's disease," *NeuroImage : Clinical*, vol. 24, p. 101954, July 2019.

[24] I. Dekker *et al.*, "The sequence of structural, functional and cognitive changes in multiple sclerosis," *NeuroImage : Clinical*, vol. 29, p. 102550, Dec. 2020.

[25] A. Eshaghi *et al.*, "Progression of regional grey matter atrophy in multiple sclerosis," *Brain*, vol. 141, pp. 1665–1677, June 2018.

[26] N. P. Oxtoby *et al.*, "Sequence of clinical and neurodegeneration events in Parkinson's disease progression," *Brain*, vol. 144, pp. 975–988, Feb. 2021.

[27] P. A. Wijeratne *et al.*, "A Multi-Study Model-Based Evaluation of the Sequence of Imaging and Clinical Biomarker Changes in Huntington's Disease," *Frontiers in Big Data*, vol. 4, p. 662200, Aug. 2021.

[28] J. L. Panman *et al.*, "Modelling the cascade of biomarker changes in GRN-related frontotemporal dementia," *Journal of Neurology, Neurosurgery & Psychiatry*, Jan. 2021.

[29] E. L. van der Ende *et al.*, "A data-driven disease progression model of fluid biomarkers in genetic frontotemporal dementia," *Brain: A Journal of Neurology*, p. awab382, Oct. 2021.

[30] M. C. Gabel *et al.*, "Evolution of white matter damage in amyotrophic lateral sclerosis," *Annals of Clinical and Translational Neurology*, vol. 7, pp. 722–732, May 2020.

[31] D. P. Kingma and M. Welling, "Auto-Encoding Variational Bayes," *arXiv:1312.6114 [cs, stat]*, May 2014. arXiv: 1312.6114.

[32] L. Antelmi *et al.*, "Sparse Multi-Channel Variational Autoencoder for the Joint Analysis of Heterogeneous Data," in *International Conference on Machine Learning*, pp. 302–311, PMLR, May 2019.

[33] Y. Xu *et al.*, "Explainable Dynamic Multimodal Variational Autoencoder for the Prediction of Patients with Suspected Central Precocious Puberty," *IEEE Journal of Biomedical and Health Informatics*, pp. 1–1, 2021.

[34] J. Cheng *et al.*, "Multimodal Disentangled Variational Autoencoder With Game Theoretic Interpretability for Glioma Grading," *IEEE Journal of Biomedical and Health Informatics*, vol. 26, pp. 673–684, Feb. 2022.

[35] D. P. Kingma *et al.*, "Semi-Supervised Learning with Deep Generative Models," *arXiv:1406.5298 [cs, stat]*, Oct. 2014. arXiv: 1406.5298.

[36] S. N, B. Paige, *et al.*, "Learning Disentangled Representations with Semi-Supervised Deep Generative Models," in *Advances in Neural Information Processing Systems*, vol. 30, Curran Associates, Inc., 2017.

[37] F. Berkhahn *et al.*, "Augmenting Variational Autoencoders with Sparse Labels: A Unified Framework for Unsupervised, Semi-(un)supervised, and Supervised Learning," *arXiv:1908.03015 [cs, stat]*, Nov. 2019. arXiv: 1908.03015.

[38] T. Ji *et al.*, "Multi-Modal Anomaly Detection for Unstructured and Uncertain Environments," in *Proceedings of the 2020 Conference on Robot Learning*, pp. 1443–1455, PMLR, Oct. 2021.

[39] V. Kmetzsch *et al.*, "A multimodal variational autoencoder for estimating progression scores from imaging and microRNA data in rare neurodegenerative diseases," in *Medical Imaging 2022: Image Processing*, vol. 12032, pp. 376–382, SPIE, Apr. 2022.

[40] T. Hastie and W. Stuetzle, "Principal Curves," *Journal of the American Statistical Association*, vol. 84, no. 406, pp. 502–516, 1989.

[41] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," *arXiv:1412.6980 [cs]*, Jan. 2017. arXiv: 1412.6980.

[42] F. Pedregosa *et al.*, "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, no. Oct, pp. 2825–2830, 2011.

[43] E. Huntzinger and E. Izaurralde, "Gene silencing by microRNAs: contributions of translational repression and mRNA decay," *Nature Reviews Genetics*, vol. 12, pp. 99–110, Feb. 2011.

[44] M. Grasso *et al.*, "Circulating miRNAs as biomarkers for neurodegenerative disorders," *Molecules (Basel, Switzerland)*, vol. 19, pp. 6891–6910, May 2014.

[45] M. D. Robinson and A. Oshlack, "A scaling normalization method for differential expression analysis of RNA-seq data," *Genome Biology*, vol. 11, p. R25, Mar. 2010.

[46] M. D. Robinson *et al.*, "edgeR: a Bioconductor package for differential expression analysis of digital gene expression data," *Bioinformatics (Oxford, England)*, vol. 26, pp. 139–140, Jan. 2010.

[47] K. Street *et al.*, "Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics," *BMC Genomics*, vol. 19, p. 477, June 2018.

[48] T. Zhou *et al.*, "Effective feature learning and fusion of multimodality data using stage-wise deep neural network for dementia diagnosis," *Human Brain Mapping*, vol. 40, pp. 1001–1016, Feb. 2019.

[49] T. Zhou *et al.*, "Hi-Net: Hybrid-Fusion Network for Multi-Modal MR Image Synthesis," *IEEE Transactions on Medical Imaging*, vol. 39, pp. 2772–2781, Sept. 2020.