# Voxelwise rs-fMRI representation learning

## A non-linear variational approach

by

E. P. T. Geenjaar

to obtain the degree of Master of Science
at the Delft University of Technology,
to be defended publicly on Monday July 12, 2021 at 16:00 PM.

| | | |
|---|---|---|
| Student number: | 4373480 | |
| Project duration: | June 1, 2020 – May 20, 2021 | |
| Prof. Dr. ir. W. Niessen | Chairman Biomedical Engineering | TU Delft, ErasmusMC |
| Prof. Dr. ir. B. P. F. Lelieveldt | Chairman Embedded Systems | TU Delft, LUMC |
| Dr. T. J. H. White, | Daily Supervisor | ErasmusMC |
| Prof. Dr. ir. A. J. van Genderen | Committee Member | TU Delft |
| Prof. Dr. ir. V. D. Calhoun | External Supervisor | TReNDS center |

# Preface

This thesis concludes my time at the Delft University of Technology for now. A time in which I had the opportunity to meet some of the most amazing people, explore myself, and have fun. I feel blessed to have been able to perform research that I feel so passionate about and have such supportive supervisors, friends, and family. Doing research and writing this thesis would have been infinitely harder without you all.

The TU Delft has been a place where I have grown both personally and professionally on my path towards this double degree. When I started out as an Industrial Design Engineering undergraduate, I had no idea that I would end up doing a double degree in Embedded Systems and Biomedical Engineering. I will always carry my time at the TU Delft with me, as I start my new chapter at the Georgia Institute of Technology and the TReNDS center to contribute to the field of neuroimaging and further our collective understanding of the brain and mental disorders.

*E. P. T. Geenjaar*
*Delft, July 2021*

# Contents

## Abstract

Resting-state fMRI (rs-fMRI) has become an important imaging modality and is commonly used to study intrinsic brain networks. These networks can be obtained by decomposing rs-fMRI data into components, using independent component analysis (ICA). Recently, these ICA components have been used as inputs for neural networks to learn complex relations between the intrinsic networks of the brain and mental disorders or demographic variables. Instead of training a non-linear classifier on these linearly decomposed components, this work asks whether unsupervised representation learning can lead to linearly separable representations for multiple downstream tasks.

We propose to apply non-linear representation learning to voxelwise rs-fMRI data. Learning the non-linear representations is done using two versions of a variational autoencoder (VAE). The first version is a vanilla VAE with 3D residual blocks in both its encoder and decoder. The second version is based on the identifiable VAE and uses a time-dependent prior. The models train to reconstruct the original input data from latent variables it infers. Three predictive models then evaluate the predictive power of the latent variables on an age regression, a sex classification, and a schizophrenia classification task. Each of the predictive models performs predictions for each of the three tasks. The predictive models are a support vector machine (SVM), a k-nearest neighbor (k-NN) model, and a long short-term memory (LSTM) neural network.

We show that our method performs exceptionally well on the age regression and sex classification tasks without any supervision. These results imply that VAEs can model predictive variations in their latent spaces for demographic variables. The models, however, do not do well on the schizophrenia classification task, even when the models are pretrained. Despite the lower performance on the schizophrenia classification task, the overall results are encouraging and pave the way for future work on voxelwise representation learning.

*Keywords: variational autoencoder, resting-state fMRI, schizophrenia, deep learning, neural networks, representation learning.*

# 1. Introduction

## 1.1. Significance

It is important to understand the functionality of the brain, not only as a philosophical pursuit, but more so to be able to understand and effectively treat mental disorders. The brain is inherently a functional dynamic system governed by the interactions between and the firing of neurons. Complex diseases like schizophrenia are considered neurodevelopmental disorders that can develop transiently [9, 35] and are related to a wide range of factors, such as genetic susceptibility, demographic variables such as paternal age [57], the external environment, and even effects down to molecular changes [9]. This is why it is of the utmost importance to both understand the development of these changes but also what changes in the brain are linked to mental disorders or the risk of obtaining mental disorders in the future. Especially since mental disorders have one of the highest mortality rates among the most substantial deaths worldwide [87].

Mental disorders are studied through multiple neuroimaging techniques, but resting-state functional MRI (rs-fMRI) has become increasingly important because it allows researchers to image the functional dynamics of a subject's resting brain over time. There has been tremendous progress in the understanding of mental disorders through linear representation learning techniques, such as independent component analysis (ICA) [15, 59]. These techniques have opened up the ability to study large-scale functional connectivity differences in complex mental disorders such as autism spectrum disorder (ASD) and schizophrenia, either statically or dynamically [12, 36, 61]. The success of linear representation learning and the increased use of deep learning methods in the field of neuroimaging [53], paves the way towards analyzing these functional differences using deep learning techniques. Findings obtained with deep learning analyses can be complemented with previous research and linear representation learning to move towards individualized predictions [80], a better understanding of mental disorders, and more effective individualized treatment.

## 1.2. Context

Deep learning methods in fields other than rs-fMRI analysis are often applied to minimally processed data. An important side note here is that these non-fMRI datasets are often also one or two magnitudes larger than many rs-fMRI datasets. In the case of rs-fMRI analysis, however, most methods use supervised classification methods. These methods are generally used because neural networks gained attention for their outstanding classification performance, most famously on ImageNet [48, 76]. The non-linearities that neural networks can model are thus interesting when it comes to mental disorder classifications too. Especially because deep learning models such as convolutional neural networks [20, 51] can more efficiently work with large input dimensionalities compared to more classical machine learning methods, for example. A 3-dimensional adaptation of supervised convolutional neural networks has been shown to obtain robust discriminative neuroimaging biomarkers [1].

Methods like independent component analysis (ICA) and principal component analysis (PCA) assume that the data is generated through some unseen factors. These factors are often referred to as intrinsic networks for ICA because they are spatially independent and interpretable as separate localized functional networks in the brain. The idea of finding the generative factors in rs-fMRI data has been extended to early unsupervised deep learning models like restricted Boltzmann machines (RBMs) [29]. Other unsupervised methods have since been popularized for non-rs-fMRI data, especially variational autoencoders (VAEs) [42] have gained attention due to their interpretable latent space and ability to variationally learn generative factors that fit a certain prior. Previous work evaluates representation learning with VAEs on rs-fMRI data that has first undergone dimensionality reduction [39, 58, 88, 89]. These dimensionality reductions may incur overly specific inductive biases and as a result limit the expressivity of deep learning methods, especially since neural networks, are considered universal function approximators [31].

### 1.3. Problem statement

This work, therefore, looks at whether unsupervised deep learning methods can learn informative representations from minimally processed voxel-wise rs-fMRI data that has not undergone dimensionality reduction(s). Similar to previous work [1] and unpublished work [2], this work evaluates representations on downstream age regression and sex prediction tasks. The downstream tasks in this work are however evaluated with representations obtained through unsupervised learning. Due to its success as an unsupervised representation learning technique, this work uses a variational autoencoder (VAE) [42], which learns to maximize the lower bound on the marginal likelihood of the training data.

One of the main reasons that this work considers the age regression and sex prediction task as the main downstream tasks, is the availability of a large rs-fMRI dataset that records both demographic factors. In this work, we find that large datasets are necessary for representation learning from minimally processed rs-fMRI data. The introduced method is also evaluated on three combined schizophrenia datasets with and without first pre-training the model on the larger age regression and sex prediction dataset [16, 56]. To evaluate the effect of the dimensionality of the representations on downstream task performance, the age and sex prediction tasks are performed with representations of varying sizes. These results are compared to a linear baseline that performs principal component analysis (PCA) with a varying number of components.

## 2. Resting-state fMRI

### 2.1. Physiology

Studying the functionality of the brain has long been an area of interest with many research fields emerging from its search for answers. Measuring what parts of the brain are activated when performing a certain task is one of the main ways in which the brain's functionality is studied. Brain activity is generally defined as the combined firing of multiple spatially coherent neurons. To find out which brain areas are active, metrics have been devised to measure the activity of neurons or groups of neurons. One common way of measuring brain activity in the 1980s was to measure the regional blood flow near the neurons in the cortex of the brain; the regional cerebral blood flow (rCBF). This was mostly done using positron emission tomography (PET) imaging which, as a drawback, has a low spatial resolution. In 1990, Ogawa, who worked at Bell Labs at the time, was the first person to show that magnetic resonance imaging (MRI) can be used to measure the rCBF with a much better spatial resolution than PET scanners [64]. They named it contrast, dependent on blood oxygenation, the finding evolved into what is now known as functional magnetic resonance imaging (fMRI) and has become one of the most popular imaging modalities to study brain activity.

### 2.2. Blood oxygenation level-dependent imaging

Magnetic resonance imaging (MRI) uses the magnetic properties of molecules to image anatomy. The contrast between oxygenated and deoxygenated blood is caused by hemoglobin. Deoxygenated hemoglobin (dHb), which is hemoglobin that has not bound any oxygen molecules, causes disturbances in the applied magnetic field in the MRI scanner [54]. These small magnetic fields are referred to as paramagnetism. Oxygenated hemoglobin (oHb) is not paramagnetic. Hemoglobin is found in red blood cells and the relative concentrations of dHb and oHb can be measured due to the field inhomogeneities that are caused by dHb's paramagnetic property [54]. Simply put, Ogawa et al. [64] have shown that the relative amount of oxygen delivery relative to the oxygen consumption can be imaged. Oxygen consumption leads to deoxygenated hemoglobin and oxygen delivery leads to larger concentrations of oxygenated hemoglobin. This imbalance is a proxy for neural activity, the blood oxygenation level-dependent (BOLD) signal can therefore be used as an approximate measure of brain activity. It is important to note that an increase in brain activation causes an increase in BOLD signal. The BOLD signal is the most common method used in fMRI imaging. A simplified representation of BOLD imaging is shown in Figure 1.

## 2.3. BOLD in rs-fMRI

The brain remains functional during rest and because the BOLD signal images the underlying physiological process, fMRI time series imaged in its resting state can lead to a better understanding of the functional connectivity of the brain. The relationship between transient blood flow and electrocortical activity [22] and the functional relationship between different brain regions [19] were already being explored before its application in fMRI. Electrocortical activity is invasive to record, however, because it requires electrodes to be placed on the brain, inside of the skull. It also only images small parts of the brain at once.

The transient blood flow patterns imaged with fMRI can be used to find functional connectivity between brain regions at rest [7]. Recording brains at rest with fMRI is called resting-state fMRI (rs-fMRI) and has become a growing and important research field. Resting-state fMRI can be used to find temporally correlated activations in separate brain regions.

The signal obtained with rs-fMRI is often assumed to be decomposable into a smaller number of generative factors. The behavior of the brain as a distributed network of these generative factors can help further our understanding of the organization and connectivity of the brain. This can in turn lead to a better understanding of brain disorders [18].

## 2.4. Schizophrenia

Schizophrenia is a mental disorder that is related to both structural and functional changes in the brain. It causes a disintegration between a person's thoughts and emotions. Patients with schizophrenia can start hallucinating, stop being able to express emotions effectively, and have difficulties with attention, concentration, and memory [63]. Finding representations that can accurately classify schizophrenia will lead to a better understanding of this incredibly complex disease, and may also lead to a more objective diagnostic tool that could aid a psychiatrist with their diagnosis. Underlying non-linear dynamics likely give rise to complex diseases like schizophrenia. Analyzing schizophrenia using non-linear representation learning is therefore important complementary information to the existing linear methods with which the disorder is often studied.

# 3. Related work

Rs-fMRI data is extremely high dimensional compared to common deep learning datasets. An rs-fMRI time series consists of several timesteps, where each timestep is a volume. This can lead to over a million voxels per subject, whereas most deep learning datasets are roughly $32 - by - 32$ or $64 - by - 64$ pixels. The most common dimensionality reduction technique that is used to study rs-fMRI data is ICA. It decomposes the 4D rs-fMRI data using a linear matrix decomposition into independent spatial components, where each component has a corresponding time series. These independent components can be mixed back into 4D data using the matrix that is learned [5, 36]. Other techniques may rely on regions of interest (ROIs) obtained using an atlas, or by preprocessing the data using software such as FreeSurfer [17].

Some interesting prior work has used variational autoencoders to analyze rs-fMRI data before. These works have focused on modeling functional brain networks and ADHD identification [72], representation learning [39], automatically clustering connectivity patterns [89], schizophrenia, bipolar disorder and autism spectrum disorder classification [58], and spatio-temporal trajectory identification [88]. It is important to note however that each of these methods performs spatial and sometimes temporal dimensionality reduction before they use the data as input for their VAE. In this work, we want to look at the applicability of VAEs for representation learning applied to voxel-wise data.

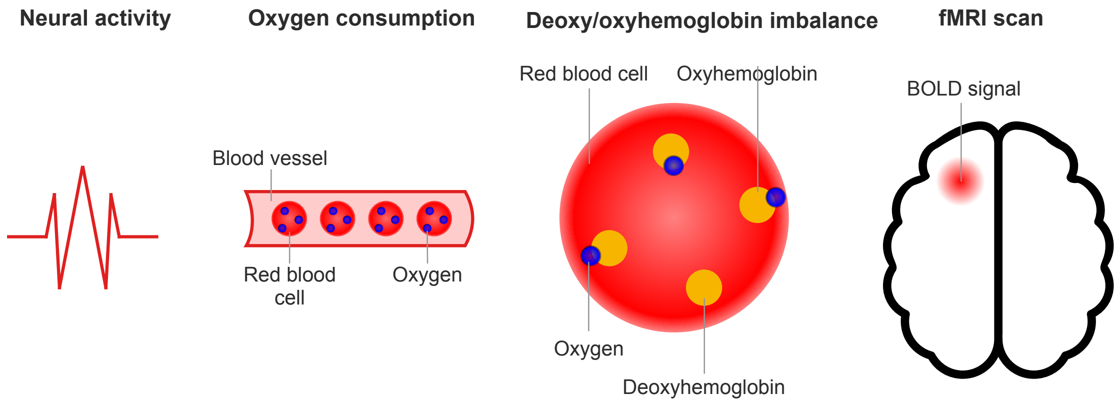


Figure 1: A simplified representation of signal that fMRI scanners pick up on, known as BOLD signal.

## 3.1. Age/sex prediction

Age and sex prediction is not of direct use to medical specialists in the way that biomarkers for mental disorders may be. Being able to predict age and sex from rs-fMRI data is valuable, however, because it furthers our knowledge about the aging of the brain and functional differences in the brain related to the spectrum of biological sex. Knowledge about functional changes in the general population due to aging may be valuable in our understanding of age-related diseases such as Alzheimer's disease. Measures like the gap between the predicted age of the brain and a person's biological age may be valuable biomarkers for various disorders.

Prior work on age and sex prediction with rs-fMRI data does so by looking at functional differences [86] or by reducing the dimensionality of the data using an atlas and treating each ROI as a node in a graph [21]. Another paper uses rs-fMRI data as a modality together with sMRI and diffusion MRI (dMRI) data to predict age [62]. Each of these methods performs spatial dimensionality reduction. Other relevant recent work includes the use of lightweight neural networks to perform age and sex prediction based on sMRI data [69]. An important difference between this work and the works mentioned above is that targets are not available to the model during training. This likely leads to lower performance compared to supervised methods, because the VAE is not specifically optimized for age regression nor sex prediction. Although achieving comparable performance to supervised methods, would indicate that unsupervised representation learning can lead to linearly separable representations. This work aims to evaluate the meaningfulness of representations that are learned in an unsupervised manner using VAEs.

## 3.2. Schizophrenia prediction

The early identification or general understanding of mental disorders is an important topic of research in the neuroimaging research community. One highly complex mental disorder that shows functional group differences [12, 36] is schizophrenia. Another paper that uses a VAE looks at semi-supervised schizophrenia classification [58] using ROIs. Other work uses the functional networks extracted from rs-fMRI data to predict schizophrenia diagnoses with 3-dimensional convolutional neural networks [73]. Both of these works require the targets to be seen during training. Other work that focuses on unsupervised representation learning for schizophrenia uses locally linear embeddings with rs-fMRI ROIs as input [78]. rs-fMRI data is also used in a multimodal study, where translation between functional and structural MRI data produces an alignment score that is correlated with schizophrenia [71].

# 4. Variational Autoencoders

VAEs have become an important model architecture for unsupervised [43] and semi-supervised learning [44]. Variational methods are a way of describing an unknown probability density function that is hard to sample from and/or approximate, byways of parametrizing a simpler probability distribution, such as a Gaussian, in such a way that it can approximate the unknown distribution [55]. Since VAEs are generative models, the goal is essentially to find the underlying latent factors that generate the data. This ties back to the assumption in methods like ICA and PCA, that the dimensionality of the latent factors is smaller than the original dimensionality of the data.

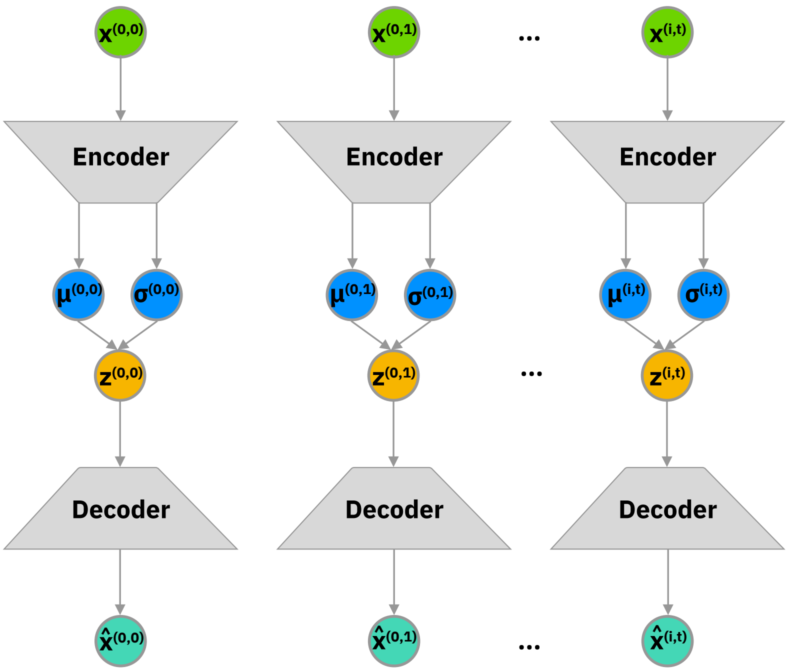


Figure 2: A visual representation of the variational autoencoder (VAE) that is used in this work. Each volume $x^{(i,t)}$ for subjects indexed by $i$ and timepoints for each subject $t$ are reconstructed $\hat{x}^{(i,t)}$ using a sample $z^{(i,t)}$ drawn from a multivariate diagonal Gaussian with mean $\mu^{(i,t)}$ and variance $\sigma^{(i,t)}$. The mean and variance are parameterized using a neural network.

The problem can formally be set as having an rs-fMRI dataset $\{X^{(i)} = \{x^{(i,0)}, ..., x^{(i,T)}\}\}_{i=1,...,N}$ with T timepoints and N subjects. Each $x^{(i,t)}$ is generated from an unknown conditional distribution $p_\theta(x^{(i,t)}|z^{(i,t)})$, where $z^{(i,t)}$ is assumed to be a random unseen continuous-valued variable sampled from a prior distribution $p_\theta(z)$ [42]. Both the prior distribution and the conditional distribution are unknown and the integral over the marginal probability of $\bar{x}$: $p_\theta(\bar{x}) = \int p_\theta(\bar{z}) p_\theta(\bar{x}|\bar{z}) \mathrm{d}\bar{z}$ is therefore intractable. Bayesian variational methods can be used to tackle this problem however and VAEs have become a common non-linear method in doing so [42].

VAEs approximate the intractable posterior distribution $p_\theta(z^{i,t}|x^{i,t})$ using a recognition model $q_\phi(z^{i,t}|x^{i,t})$ [42], which can be thought of as an encoder from a coding theory perspective. The conditional distribution $p_\theta(x^{i,t}|z^{i,t})$ can then be thought of as a decoder. This allows us to optimize the evidence lower bound (ELBO), which is a lower bound on the marginal likelihood of data point $x^{(i,t)}$:

$$\begin{aligned}\log p_\theta(x^{(i,t)}) &\geq \mathcal{L}(\theta, \phi; x^{(i,t)}) \\ &= \mathbb{E}_{q_\phi(z^{(i,t)}|x^{(i,t)}}[-\log q_\phi(z^{(i,t)}|x^{(i,t)}) \\ &\quad + \log p_\theta(x^{(i,t)}, z^{(i,t)})] \\ &= -D_{\mathrm{KL}}(q_\phi(z^{(i,t)}|x^{(i,t)}) \mid\mid p_\theta(z^{(i,t)})) \\ &\quad + \mathbb{E}_{q_\phi(z^{(i,t)}|x^{(i,t)})}\left[\log p_\theta(x^{(i,t)}|z^{(i,t)})\right]\end{aligned} \tag{1}$$

The proof is explained in detail in Kingma and Welling [42]. The objective function, with which the VAE is trained, is the average over all the data points. Both the encoder $q_\phi(z^{(i,t)}|x^{(i,t)})$ and decoder $p_\theta x^{(i,t)}|z^{(i,t)}$ are parametrized as neural networks. The loss can then be split into two parts, the first part minimizes the KL-divergence between an apriori selected prior and the distribution that is parametrized by the encoder. Although there is no clear consensus, rs-fMRI data is often seen or assumed to be normally distributed [50]. The prior ($p_\theta(z)$) we thus assume is a diagonal multivariate normal distribution, where each of the dimensions of the normal distribution do not explicitly depend on each other. The second part of Equation 1 maximizes the log-likelihood of a data point $x^{(i,t)}$ given an estimated latent variable $z^{(i,t)}$. Although each rs-fMRI time point is assumed to be an i.i.d. sample when training the VAE, the temporal relation between the latent variables for a single subject ($z^{(i,0)}, ..., z^{(i,T)}$) is considered during each prediction task. A visual representation of the VAE is shown in Figure 2.

# 5. Contributions

This work attempts to learn meaningful representations by applying a variational autoencoder to voxelwise rs-fMRI data. Deep learning has started to become prevalent in the neuroimaging community and can capture robust representations [1]. Voxelwise representation learning has especially become common for structural MRI scans, which is likely because sMRI is more readily interpretable and less noisy in comparison to rs-fMRI. Most deep learning methods that are applied to rs-fMRI data use dimensionality reductions like ROIs or ICA components. In theory, however, because neural networks are universal function approximators [31], they should be possible to learn non-linear generative factors that explain the distribution of the data even more accurately. These generative factors can lead to new insights into how our brain works functionally. This work uses variational autoencoders as a non-linear feature extraction method and evaluates its representations on several different classification tasks with multiple types of classifiers. Given that rs-fMRI data is noisy and that there is relatively little data available compared to other deep learning domains, it makes sense to pursue an unsupervised method [16] that is easily extendable and can lead the way towards non-linear discoveries. Previous work on transfer learning in rs-fMRI shows promising results [56]. Given that datasets may be small for rs-fMRI studies, this work also looks at the effect of pretraining for voxelwise representation learning using a VAE.

# 6. Data

This work uses multiple datasets, one dataset to do age regression and sex prediction with and another combination of multiple datasets to perform a schizophrenia classification task on.

## 6.1. Age and sex dataset

The rs-fMRI that is used for age and sex prediction are subjects without any diagnoses or self-reported illnesses ($n = 12,314$). These subjects were selected from the 22,392 subjects that were available in the UK Biobank repository on April 7th, 2019 [60]. The subjects have a mean age of $62.58$ with a standard deviation of $7.41$, $49.6\%$ are female. The youngest subject is 45 years old and the oldest is 80. The scanning parameters are explained in greater depth in Miller et al. [60], however, an important parameter in this work is the repetition time (TR = $0.735$ seconds). With an acquisition time of $6$ minutes, the UK Biobank data acquires a total of 490 time points. The data is minimally preprocessed with the Melodic pipeline [60] and registered to the MNI EPI template with the help of FMRIB's Linear Image Registration Tool (FLIRT). The registration is followed by normalization in SPM12, after which it is smoothed with a $6$mm wide FWHM Gaussian kernel. This results in rs-fMRI volumes with a size of $53$ x $63$ x $52$ voxels, and $490$ timepoints per subject. The size of the volumes and the number of timepoints lead to large memory requirements during training and rs-fMRI data can be noisy. To tackle both problems simultaneously, we use a piecewise aggregate approximation (PAA) to reduce the noise and memory consumption, while still keeping the trend of the time series. PAA takes the average over points in consecutive windows with a certain window size. For the UK Biobank dataset, the window size is set to $15$, which is equivalent to taking the average over a period of about $11$ seconds. This reduces the number of time points to $33$. A visual representation of PAA is shown in Figure 3.

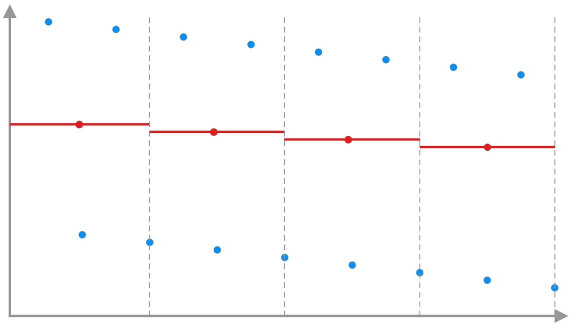


Figure 3: An example of piecewise aggregate approximation (PAA), where the blue points are the original 'noisy' points and the red points are the new points after PAA. PAA is able to keep the same general trajectory, with less noise. It creates multiple windows with a certain window size and averages the points inside those windows.

## 6.2. Schizophrenia datasets

The datasets that are used to evaluate the potential of the VAE representations to distinguish between healthy controls and patients with schizophrenia are FBIRN [37], B-SNIP [81], and COBRE [3]. Each dataset was processed using NeuroMark preprocessing pipeline [15] to obtain rs-fMRI volumes with a size of $53$ x $63$ x $52$ voxels. The number of timesteps for each dataset differs, but the repetition time is the same (TR =

$2.0$ seconds). To make sure each subject has the same number of time points, we use the lowest number of timesteps available in a dataset, which is $100$ timesteps in B-SNIP. For scans with more timesteps, we only use the first $100$ timesteps. To stay in line with the temporal preprocessing that is done for the UK Biobank dataset, we apply PAA to these datasets as well, but to account for the different repetition times, the window size that is used is $5$. This corresponds to a period of $10$ seconds.

# 7. Methodology

This work proposes to use a variational autoencoder to learn representations from voxelwise rs-fMRI time series. The model is trained on multiple datasets to evaluate its potential on noisy and complex neuroimaging data. Learning representations with non-linear methods from voxels directly should theoretically lead to better results compared to linear methods like PCA, although in practice it is hard to train large deep learning models on small datasets effectively. This is especially true for rs-fMRI data, which is also highly noisy, due to physiological noise and artifacts introduced by head movements.

The question this work tries to answer is whether the generative factors that the VAE finds, contain information about demographic variables and schizophrenia. Given that unsupervised learning is less likely to lead to overfitting on smaller datasets [16] and that VAEs have become an important method for representation learning [83], the unsupervised representations it extracts from voxelwise rs-fMRI time series may be valuable to further neuroimaging research. Another important consideration in this work is that VAEs are inherently more insightful because group differences can be decoded back into brain space. This allows for a model that does not immediately require explainability methods to get an insight into regions that may differ between groups. This reduces bias because explainability methods may have to be tuned and the decoder more accurately reflects the representations that are learned.

The performance of the representations extracted from the rs-fMRI time series by the VAE will be compared to a baseline linear method on the downstream tasks. There is no readily available baseline, however, because there is not much previous work that looks at voxelwise rs-fMRI representation learning, although unpublished work focuses on supervised sex classification and age regression on the same dataset used in this work [2]. Supervised methods, however, likely outperform unsupervised methods because they are trained specifically for the downstream tasks. Supervised methods are therefore not a comparable baseline, although comparing performance can be insightful. One of the most commonly used unsupervised methods to find the generative factors that underly rs-fMRI data is ICA. An important difference between ICA and a VAE is that when ICA is applied to rs-fMRI data, the independence constraint is enforced on the voxels in the brain space. This is not a constraint that can easily or logically be extended to VAEs. Further, ICA is generally run multiple times and spatial components may be manually selected, this also does not transfer well. Generative VAE factors look nothing like the spatially localized ICA components that are generally used in fMRI analyses, especially because the generative factors model variance throughout the brain and are not localized to specific areas. This opens up the opportunity to take a different direction and look at generative factors that may not be as visually insightful as spatial ICA components and can potentially provide complementary information about the same dataset. Furthermore, running ICA analyses on large datasets like UK Biobank, such as NeuroMark [15], with multiple numbers of components to compare the method with the VAE experiments, is too computationally expensive for this project.

Another method that linearly decomposes data into several pre-specified components is principal component analysis (PCA). It turns out that a VAE is in some sense comparable to PCA in terms of what components it tends to learn [11, 75]. Given that PCA and VAE are comparable and that there are online versions of PCA [4] that do not have large memory requirements, the baseline for this work is linear IncrementalPCA [4, 68].

Since one dataset is significantly smaller and is prone to overfitting, this work also evaluates whether a model that is pre-trained on a large dataset like UK Biobank ($\mathrm{n}=12,314$) can be fine-tuned on the joint schizophrenia datasets ($\mathrm{n}=901$), to improve representations for schizophrenia classification.

## 7.1. Model architecture

The architecture of the model is based on a ResNet [26]. Each residual block in the encoder and decoder has a skip connection, the residual block used in the encoder and decoder are shown in Figure 4. These skip connections allow the network to learn longer dependencies and have been used in VAEs before [45] to improve their variational inference. The specific skip connection in this work does not have a stochastic and deterministic path, because it does not use hierarchical latent variables. The reason it does not use hierarchically conditioned latent variables is that we assume a single hierarchy of generative factors. The skip connections in this work thus only have a deterministic path. The activations that are used in the network are exponential linear units (ELUs) [10]. Instead of batch normalization, the network uses weight normalization [77]. Both the activation function and the weight normalization are considered best practices when training VAEs [77]. Batch normalization may lead to drift during inference in a VAE which can cause unstable results.

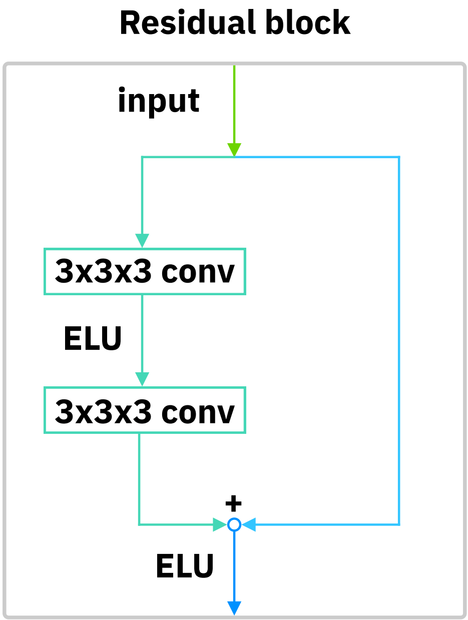


Figure 4: The residual block used in the encoder and the decoder. For the encoder, the '3x3x3 conv' is a convolutional layer with a kernel size of 3, a stride of 2, and padding of 1. This is the same for the decoder, except it uses a transposed convolutional layer and the stride is set to 1. The decoder uses trilinear upsampling layers to increase the spatial size of the activations.

The encoder consists of five residual blocks, with 16, 32, 64, 128, and 256 output channels for each block, respectively. These blocks all downscale their original inputs by two until the last residual block produces an output feature map of 256 x 2 x 2 x 2, which is flattened to 2048 features. These features are mapped in two separate linear layers to the mean and the square root of the natural logarithm of the variance for the multivariate Gaussian. The layer computes the square root of the natural logarithm of the variance instead of the standard deviation to increase the stability of training the network and to make sure variations due to gradient updates have a smaller effect on standard deviations near zero. This allows the network to model the standard deviations near zero more accurately. The first layer in the decoder is a linear layer that maps latent variable $z$ to 2048 features, which are then reshaped to a 256 x 2 x 2 x 2 feature map. We use trilinear interpolations on the feature maps with a scale of two to double the size. The rest of the decoder consists of five residual blocks, where each residual block is preceded by a trilinear interpolation layer. The final residual block is followed by a 1x1x1 transpose convolutional layer, with a stride of 1, this layer is not followed by any activation function. In earlier iterations of this work we tried to use a sigmoid activation on the last layer, this leads to non-convergence, because the combination of the mean squared error (MSE) as a loss function and a sigmoid leads to a non-convex objective function. Each of the layers is initialized according to He et al. [26], this initialization is also used in the original ELU paper [10].

## 7.2. Unsupervised training

The likelihood function that we use for the output of the decoder is a normal distribution. Directly optimizing normal distributions at each voxel turned out to be highly unstable. Maximizing a normally distributed likelihood is equivalent to minimizing the MSE between the reconstruction and the original sample, under the assumption that the standard deviation of all the voxels is the same. Replacing the normal distributions by a single point and minimizing the MSE norm turned out to be much more stable. The VAE is trained for 100 epochs using the ADAM optimizer [40] with a learning rate of $5E-4$. Before the input data is used it is first rescaled to be between [0, 1], values below $0.05$ are then thresholded to remove possible background noise.

## 7.3. Identifiable variational autoencoder

Recent work has unified previous work on nonlinear ICA [32–34] and variational autoencoders [38]. This work proves that conditioning the latent representation $z$ on an additional auxiliary variable $p_\theta(z^{(i,t)}|x^{(i,t)},u)$ leads to an identifiable model [38]. In the case of our work we choose the auxiliary variable to be the current time step of each volume: $p_\theta(z^{(i,t)}|x^{(i,t)},t)$. A visual representation of the identifiable VAE is shown in Figure 5. The multilayer perceptron (MLP) consists of three linear layers with a hidden size of $512$. The nonlinearities between the layers is an ELU [10] and each layer is weight normalized [77]. The output of the last layer in the MLP is mapped to a mean $\mu^{(t)}$ and a variance $\sigma^{(t)}$ using two linear layers, similar

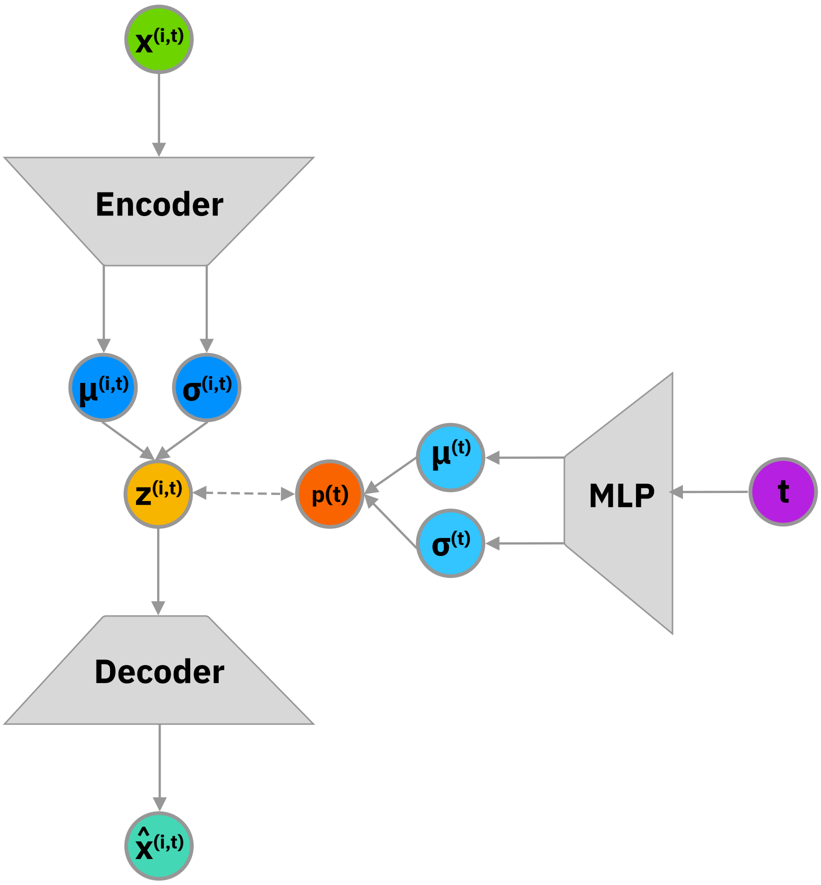


Figure 5: A visual representation of the identifiable VAE, where the prior is conditioned on the timestep $t$ of volume $x^{i,t}$. The function that is used to map $t$ to $\mu^t$ and $\sigma^t$ is a multilayer perceptron (MLP) followed by two linear layers, one for the mean and one for the variance.

to the VAE. The KL-divergence between a static prior and the predicted distribution $p_\theta(z^{(i,t)}|x^{(i,t)})$ can then no longer be used. To optimize the MLP and the VAE conjointly we instead minimize the log probability of $p_\theta(z^{(i,t)}|x^{(i,t)})$ and maximize the log probability of $p_\rho(z^{(i,t)}|t)$. This is similar to having the mean and variance of the prior be timestep-dependent and maximizing the similarity between the predicted and the timestep-dependent prior.

The identifiable VAE is compared to the normal VAE for most experiments to evaluate whether moving towards identifiable VAEs is a valuable step for future work. The identifiable VAE is henceforth referred to as an iVAE.

## 7.4. Regression and classification

After training the VAE, there are multiple ways to evaluate what information is contained in the representations $(z^{(i,0)}, ..., z^{(i,T)})$. To evaluate whether the temporal information improves classification and regression with simple machine learning classifiers, these classifiers are trained with a subject's latent temporal average $z^{(i,\mu)}$ and also with a subject's concatenated latent time series. This allows us to determine whether the temporal dynamics captured in the latent time series can be used to improve the performance on the downstream tasks. The machine learning classifiers that are used in this work are a support vector machine (SVM) and a k-nearest neighbor classifier (KNN-C) for the classification tasks and a support vector regression (SVR) and k-nearest neighbor regressor (KNN-R) for the age regression task. These classifiers give us insight into the linear separability of the representations (SVM) or how well they are clustered (KNN-C). To take the temporal information between the representations into account more specifically, we also train a long-short term memory (LSTM) [30] on the full latent time series. LSTMs have commonly been used to model time series and can model temporal relations between the points to predict a certain target variable [52]. The LSTM is either trained with a mean squared error (MSE) for the regression task or a binary cross-entropy (BCE) loss for the classification task. The hidden size for the hidden states in the LSTM ($h^{(0)}, ..., h^{(T)}$) are twice the size of the input representations, and all of the hidden states in the LSTM are concatenated together to form a feature vector that is then mapped to a prediction using a linear layer. The LSTM that is used for classification is shown in Figure 6. Since the size of that feature vector can be quite big, we apply dropout to the last layer, which is a common technique to counter overfitting and promote a more robust prediction model [79].

### 7.5. Evaluation measures

To be able to compare the results obtained using each of the methods, we use multiple evaluation measures. The first measure is used for the classification tasks and computes the area under the receiver operating characteristic (ROC-AUC), this is a more complete way of comparing binary classifiers. The ROC-AUC does not only look at the number of correctly classified samples, but also takes into account the false positive rate (FPR), and the true positive rate (TPR). To evaluate the regression task, we use three measures, the first is the mean average error (MAE) which is the L1-norm between the predicted age and the correct age. The second measure is the R2-score, which is also referred to as the coefficient of determination. It is proportional to the explained variance in the true variable. The last measure is the Pearson product-moment correlation between the predicted ages and the true ages. We use multiple measures to obtain a more complete picture of the predictions because a low MAE may, for example, not imply that the predictions necessarily explain the variance in the true variables.

## 8. Experiments

### 8.1. Experimental settings

The code for the VAE was implemented using PyTorch [67], training was performed with Catalyst [47] and TorchIO [70], and the regression and classification pipelines were implemented using RAPIDS-AI [82], scikit-learn [68], and NumPy [85]. Most if not all code was written by the authors of the paper, especially because voxelwise rs-fMRI data requires many custom pipelines due to the size of the data. To minimize costly transfers between the CPU and the GPU, most of the classifications were done using RAPIDS-AI [82] , to make sure the computed representations could be kept in GPU memory without any copies or transfers from or to the CPU. All of the experiments were performed on an NVIDIA DGX-1 V100. Due to time restrictions, the UK Biobank experiments could only be performed on one train and test split, because each VAE epoch takes around 45 minutes on a single GPU. Training on multiple training and test folds is essential for the schizophrenia task because the variance between predictions can be large, especially for deep learning models. The schizophrenia results are thus trained over 5-folds, where one fold is used as the held-out set and the other four folds are used as a training and validation set. To make sure the model does not overfit, we use an early stopping criterion that stops the model if its loss objective has not improved on the validation set for 20 epochs.

### 8.2. Latent dimensionality

To determine the effect that the size of the representations has on the performance of the age regression and sex classification tasks, the model is trained with multiple latent dimensionalities, specifically: 64, 128, 256, and 512. These tests are done on the UK Biobank dataset because we noticed that training on a larger dataset is more stable.

### 8.3. Pre-training

Pre-training is valuable for rs-fMRI [56] and as regularization for low-data regimes in deep learning [16]. To do so, we train on a larger dataset and finetune the weights of the neural network on a smaller dataset. In this case, the work contains two datasets, one large dataset: UK Biobank, and one smaller dataset that is comprised of multiple smaller datasets: FBIRN, COBRE, and BSNIP. Given that the UK Biobank dataset is about 12 times bigger, we tested whether initializing a model for the schizophrenia classification task with a pre-trained model on UK Biobank improves the results on that task. The number of latent dimensions that are used for the transfer learning task is $256$.

### 8.4. Baseline

The baseline for this work is IncrementalPCA, which is implemented in scikit-learn [68]. It is a PCA that can be trained using batches, which is necessary because the dataset can not easily be kept in memory due to its size. Just like with the VAE, the principal components are obtained for each volume in a subject's time series independently. The components are also whitened and after finding the components, they are averaged over time and then used as input in an SVM and an SVR for the sex classification and age regression

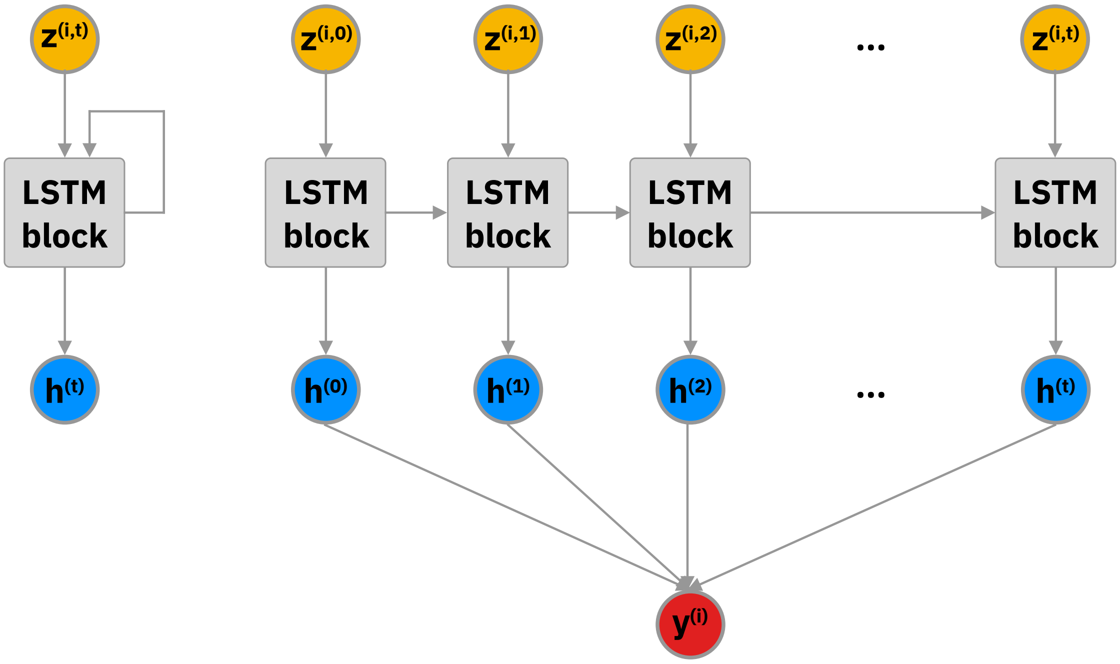


Figure 6: A visual representation of the way the LSTM is used to come to a prediction for a subject. The hidden state for each time point is taken into account for the prediction. The representation on the right is equivalent to the unrolled version on the right.

tasks. The sex classification and age regression tasks are used because they can be done on a larger dataset, which leads to more robust results. The schizophrenia task does not always converge well for the VAE.

# 9. Results

## 9.1. Latent spaces

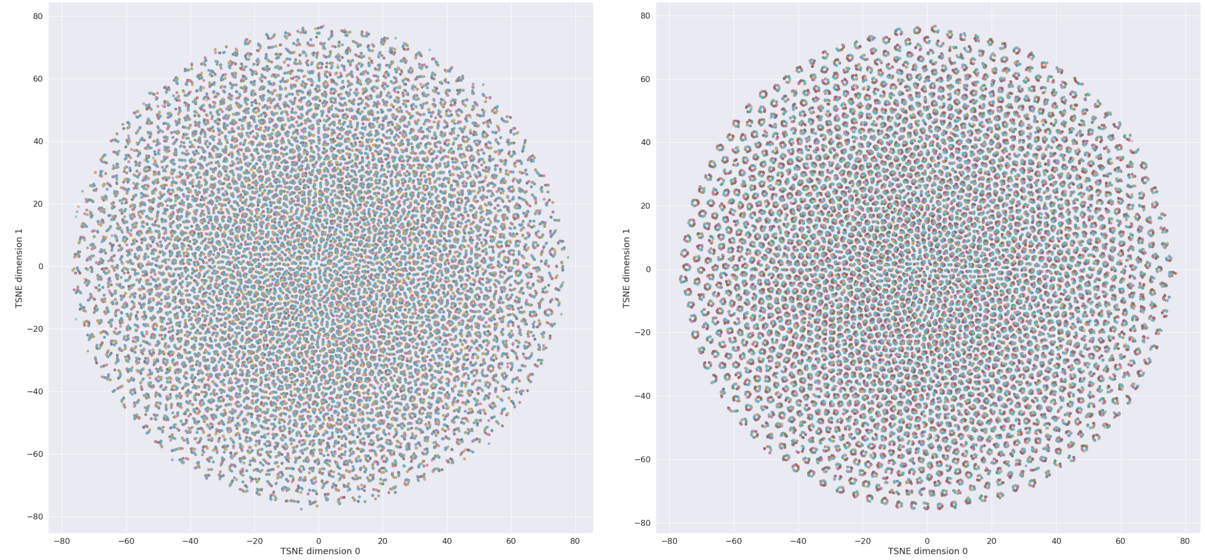
The iVAE is trained with a modified objective, such that the location of the prior for a volume is based on an auxiliary variable, in this case, the timestep of a volume. To inspect how the representations for an iVAE differ from a VAE in the test set, the representations are visualized as a 2D figure using t-SNE [84], as seen in Figure 7. t-SNE approximately preserves local and global distances for high-dimensional points in a 2-dimensional plot. The perplexity, a t-SNE hyperparameter, is set to a low value for this plot to focus on the local differences. Each timestep is assigned a different color. Although the plots are fairly similar, the timesteps in Figure 7b for the iVAE seem to be more dispersed around a subject's cluster. The increased spread of the timesteps in the latent space is likely caused by the iVAE's flexible timestep prior. Furthermore, in both Figures 7a and 7b, the subjects seem to be clustered in the latent space. This points towards VAEs learning timesteps from the same subjects as being similar. One reason that may cause this is the way the batches are constructed. The batches consist of $4$ subjects for this task and because each subject has $33$ timesteps, the optimizer takes gradient steps based only on a few subjects and more so based on the timesteps for each subject. This may bias the optimizer towards finding local minima that cluster subjects.

## 9.2. Latent dimensionality

The age and sex downstream tasks are evaluated for multiple latent dimensionalities and compared with a baseline PCA that has the same number of components. The classification methods are referred to as SVM and kNN when the representations for each timestep are concatenated to create a single feature vector ($z^{(i,0)}$, ..., $z^{(i,t)}$) for each subject. Classifiers mSVM and mkNN take the average representation over the timesteps as input for each subject. The $128$-dimensional VAE did not converge well which leads to worse results for that specific model.

### Age regression results

The age regression task is evaluated using three measures, the mean absolute error (MAE) is seen as most important in this work, the results for the task are shown in Figure 8. All of



(a) A t-SNE plot for a 256-dimensional VAE

(b) A t-SNE plot for a 256-dimensional iVAE

Figure 7: These are two t-SNE plots with low perplexity to show the local differences in the latent space on the test set for both a VAE (left) and an iVAE (right). It seems like the subjects with their timesteps, each timestep in a different color, are clustered. Noticeably, the timesteps in the iVAE TSNE plot are further apart, which is likely caused by the iVAE's flexible time prior.

the VAE models, even the non-converging $128$-dimensional VAE, outperform the baseline PCA method. All of the VAE models, except the $128$-dimensional VAE, outperform the iVAE. The best performing model is the $512$-dimensional VAE-SVM with an MAE of $4.014$ years, an R2 score of $0.5288$, and a correlation between the predicted and ground truth ages of $0.727$. The general trend for the number of latent dimensions is that more latent dimensions improve the downstream performance on the age regression task. The difference between the $256$-dimensional VAE and the $512$-dimensional VAE is significantly smaller than between the two smaller latent dimensionalities. Furthermore, the SVM performs roughly on par with the mSVM for the VAE and iVAE although sometimes slightly better, this seems to align with the clustering of subjects in Figure 7. If the subjects are clustered with their timepoints, the timepoints will likely not contain any extra information for the regression nor classification task. Another interesting result is that the SVM and mSVM always outperform the kNN and mkNN, which suggests that the latent space is linearly separable for the age regression task, as opposed to clustered based on age. Furthermore, the SVM and mSVM also outperform the LSTM, which makes sense if the subject clustering hypothesis is true. The LSTM highly focuses on temporal information and it may struggle with data where temporal relations do not aid in regression improvements. The LSTM surprisingly does perform well for the correlation between the predicted and ground truth ages.

Comparatively, unpublished work [2] reports that on the same dataset, a voxelwise supervised deep learning model achieves an MAE of $3.54$ years, an R2 score of $0.65$, and a correlation of $0.82$. It is important to note that the VAE model in this work received no supervised signal to model the features necessary to achieve its downstream age regression results. Further, the unpublished work [2] finds that using the ICA time series achieves an MAE of $4.66$, which is worse than the results reported in this work. Further, they find that taking the mean over the temporal dimension before using the input for the classification task improves performance, which seems to be the same in this work.

To visualize how age is distributed throughout the latent space, representations for the training set, validation set, and test set are encoded using the $256$-dimensional VAE and visualized in a t-SNE plot [84] in Figure 9b. Although there is no clear age gradient in the t-SNE plot between younger and older subjects, it seems like the older subjects are in general clustered more towards the bottom right part of the plot, with the younger subjects in the top left and middle left. The weak latent space gradient between ages in the t-SNE plot does not mean that that structure does not exist in the $256$-dimensional latent space. In fact, given the good results on the age regression task, the structure likely exists but is hard to visualize in a

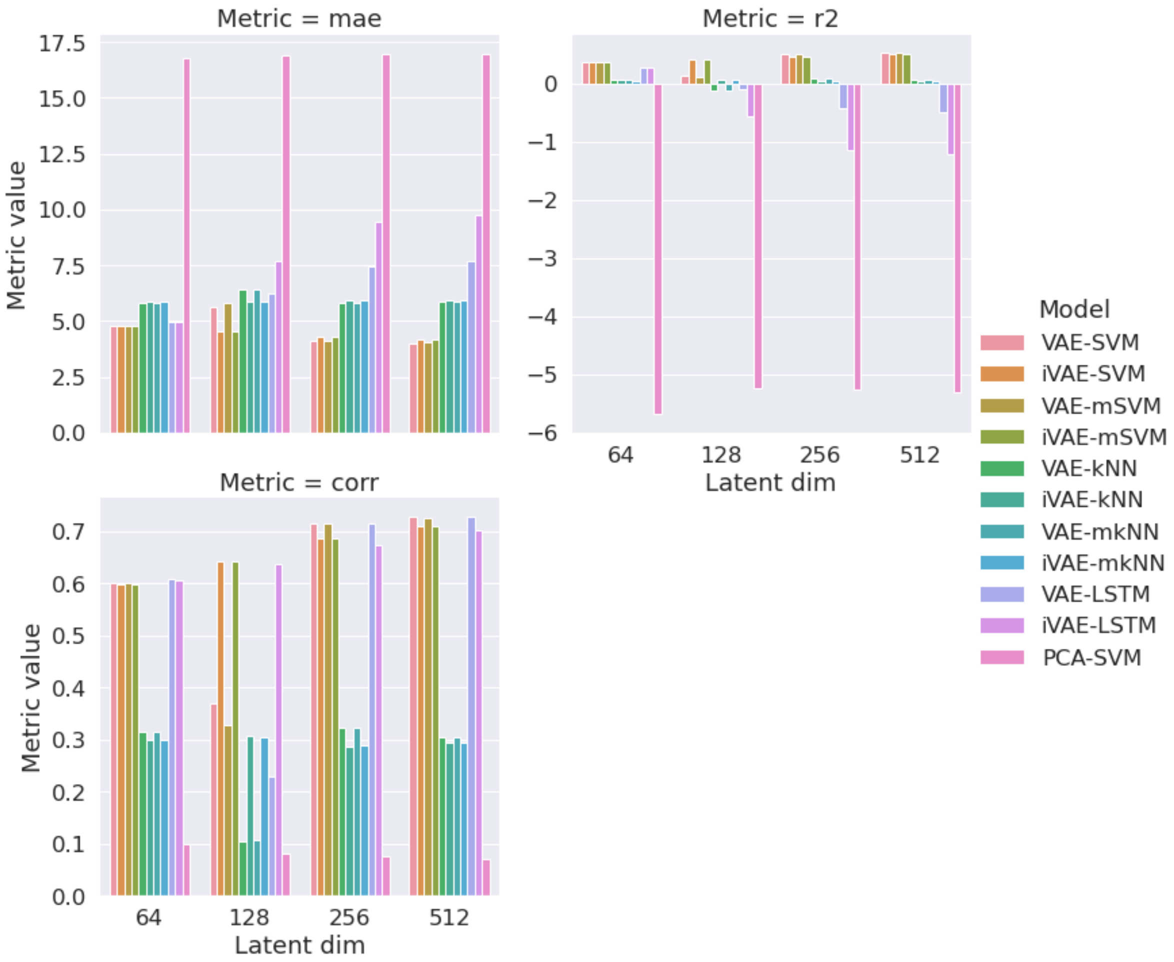


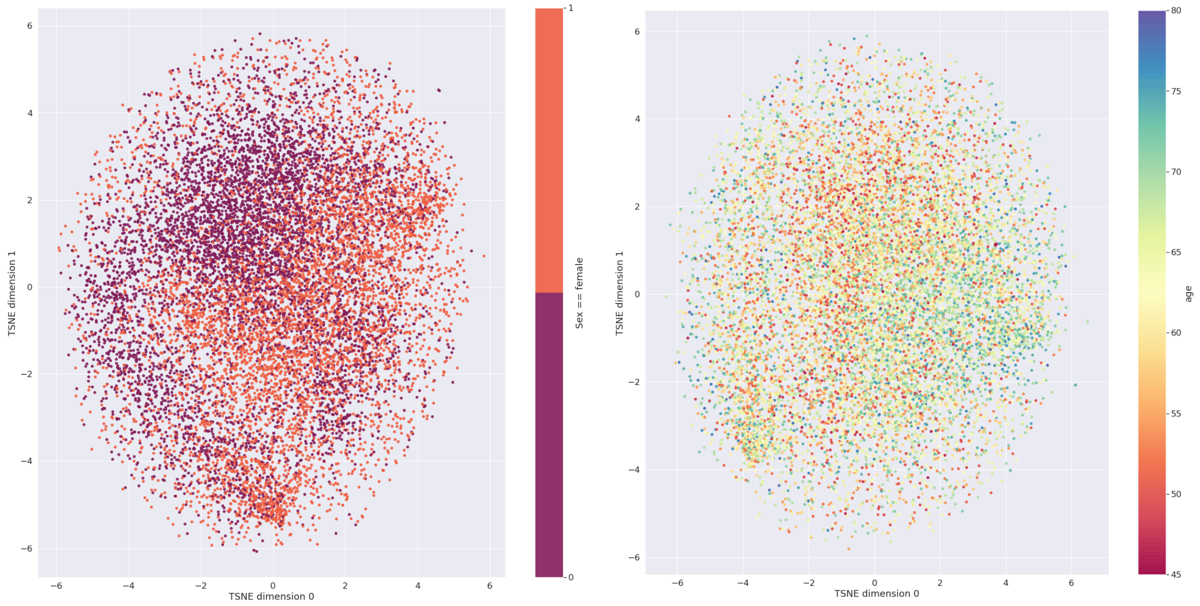
Figure 8: The results for the downstream age regression task, the three metrics are MAE (mean absolute error), the R2 score, and the correlation between the predictions and the ground truth values. Each bar plot shows all of the models at the $4$ different latent dimensionalities: $64$, $128$, $256$, $512$ on the x-axis. The run with a $128$-dimensional VAE did not converge well, an anomaly, which leads to worse results but, a VAE-SVM performs best overall. Note that for the MAE lower is better, and for the R2 score and correlation higher is better. The baseline performs significantly worse on all three of the metrics. Further, the VAE and iVAE-LSTM perform relatively well on the correlation metrics but significantly underperform on the MAE and the R2 score.

2-dimensional image.

**Sex classification results**

The sex classification task is evaluated using the area under the curve for the receiver operating characteristic (ROC-AUC). As opposed to evaluating the task using classification accuracy, ROC-AUC also takes the false-positive rate and false-negative rate into account. The results show that the age classification task for this dataset is fairly trivial and the baseline performs only slightly worse than the VAE-SVM and VAE-mSVM. The baseline outperforms the $128$-dimensional VAE though because that model did not converge. In general the baseline also slightly outperforms all forms of the iVAE, which is surprising, but only underlines that the combined effect of a flexible prior and subject clustering may hurt performance. The root cause of the problem is the subjects clustering in the latent space because it may reduce the group-wise features that are modeled.

All of the models improve with increasing latent dimensionality, although the models only slightly improve after $256$ dimensions. Interestingly, as opposed to the age regression results, the LSTM performs only slightly worse than the mSVM and SVM for both the VAE and the iVAE. The mkNN and kNN are always outperformed by the mSVM, SVM, and LSTM however for both the VAE and iVAE. The best performing model is the $512$-dimensional VAE-mSVM with a ROC-AUC of $0.994$. In general, except for the $128$-dimensional VAE, the mSVM outperforms the SVM, which suggests that the temporal information for each subject does not help with the linear separability of sex in the latent space. Although the SVM slightly out-



(a) A t-SNE plot for a $256$-dimensional VAE with sex colored differently

(b) A t-SNE plot for a $256$-dimensional VAE with age colored differently

Figure 9: Two t-SNE plots showing a 2-dimensional projection of the latent space learnt by a $256$-dimensional VAE. Representations in the training, validation, and test set are all included, they are combined and each subject is represented as the average representation over their timesteps. The plot where sex is colored differently shows that the representations differ in location based on sex, which is also clear from the classification results. Higher ages are slightly more clustered towards the bottom right with lower ages more clustered towards the top left.

performed the mSVM for most latent dimensionalities, the conclusion was generally the same, the temporal information does not add much to the task performance, likely because of the subject-wise clustering.

The organization of the latent space, colored based on sex, as encoded by the $256$-dimensional VAE is shown in Figure 9a. The points that are embedded are averages over the temporal dimension and come from the training, validation, and test set. The linear separability seems to be almost perfect in the $256$-dimensional space, based on the ROC-AUC reported in Figure 10. The 2-dimensional t-SNE projection seems less linearly separable, but the male subjects are embedded more on the top left side of the plot and the female subjects are embedded on the bottom right side of the plot. This shows that the VAE learns to represent sex differently without any supervision.

## 9.3. Schizophrenia classification results

The downstream schizophrenia classification task is evaluated for $3$ different models, a randomly initialized (as described in subsection 7.1) $256$-dimensional VAE, a $256$-dimensional iVAE, and a $256$-dimensional VAE that was pre-trained on the UKBB dataset. The results are visualized in Figure 11. The results are averaged over $5$ folds and the best average result is achieved by the pre-trained VAE-mSVM with a maximum ROC-AUC of $0.5914$. These results indicate that the model performs slightly better than predicting the labels randomly, which is not a convincing result. Figure 11 does show that pre-training improves the result for the VAE-mSVM and the VAE-LSTM, further, the iVAE-mSVM, iVAE-SVM, and iVAE-kNN outperform their VAE counterparts. This indicates that the flexible time prior and pre-training may be valuable to obtain better representations for a downstream schizophrenia prediction task, but that further work is needed. The mSVM also outperforms the SVM for all models, this again suggests that the temporal dimension adds no valuable information for linear separability in the latent space.

A reason for the marginal improvement with pre-training may be tied back to the subject-wise clustering in the latent space that is apparent in Figure 7a. In the case of subject-wise clustering, finetuning on another dataset will only add new subject-wise clusters but may not necessarily lead to a better local minimum with more valuable and regularized features.

## 9.4. Sex-based group differences voxel space

A VAE is a generative model, as explained in section 4 and is thus capable of reconstructing locations from the latent space. To visualize the differ-

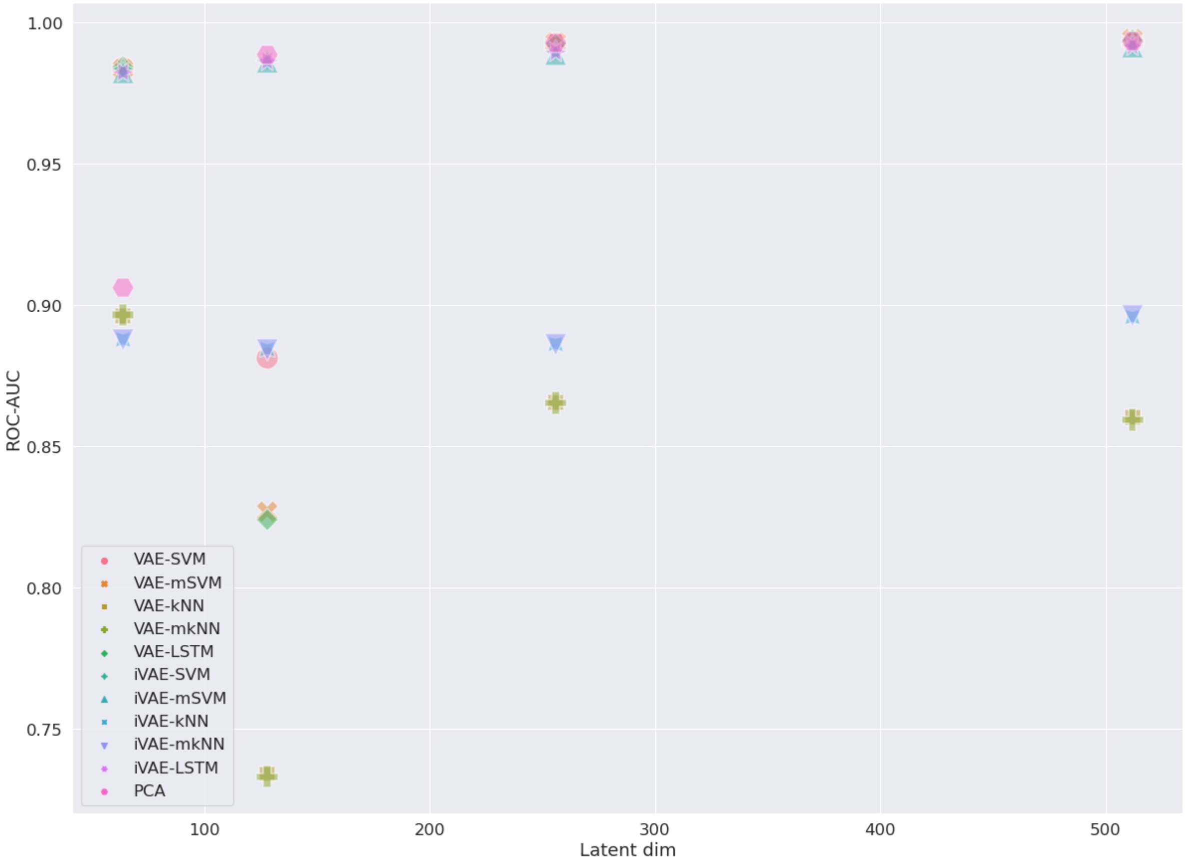


Figure 10: The results for the downstream sex classification task, the area under the curve for the receiver operating characteristic (ROC-AUC) is shown on the y-axis. The $4$ different latent dimensionalities are shown on the x-axis: $64$, $128$, $256$, and $512$. The run with a $128$-dimensional VAE did not converge well, an anomaly, but a VAE-mSVM performs best overall. The models perform similarly to the baseline, the VAE-mSVM and VAE-SVM perform slightly better, the iVAE-mSVM and iVAE-SVM perform slightly worse. The highest ROC-AUC is achieved using a $512$-dimensional VAE-mSVM: $0.994$

ences in the VAE models between males and females in its latent space, the average representation for both groups is decoded back into the voxel space and the reconstruction for males is subtracted from the reconstruction for females. The resulting volume is then thresholded at the highest 80th quantile absolute value and the differences are shown in Figure 12. The decoded group-wise differences show that women on average have increased activation in a large area of the prefrontal cortex, which has been reported in the literature before [28]. There also appears to be some increased average activity in the left and right inferior parietal lobules. The activity in the inferior parietal lobules looks more like noise and does not persist when higher thresholds are used. There are also some differences between the occipital lobe and the cerebellum outside of the brain, these differences are also likely reconstruction noise.

## 10. Conclusion

This work investigated whether unsupervised deep learning techniques, more specifically a VAE, can learn robust representations that can be used in downstream neuroimaging tasks from minimally preprocessed voxelwise rs-fMRI data. The representations learned by a VAE and an iVAE were evaluated on multiple downstream tasks and for multiple different latent dimensionalities on two of those downstream tasks. The most important downstream tasks in this work are the age regression and sex classification tasks on the UK Biobank dataset. Further, the models performed a downstream schizophrenia classification task on a combined schizophrenia dataset, which consists of B-SNIP, FBIRN, and COBRE. The downstream task was performed using a normal VAE, an iVAE, and a pre-trained VAE. Finally, the group differences for the sex classification task were visualized to understand what sex differences were learned by the VAE model.

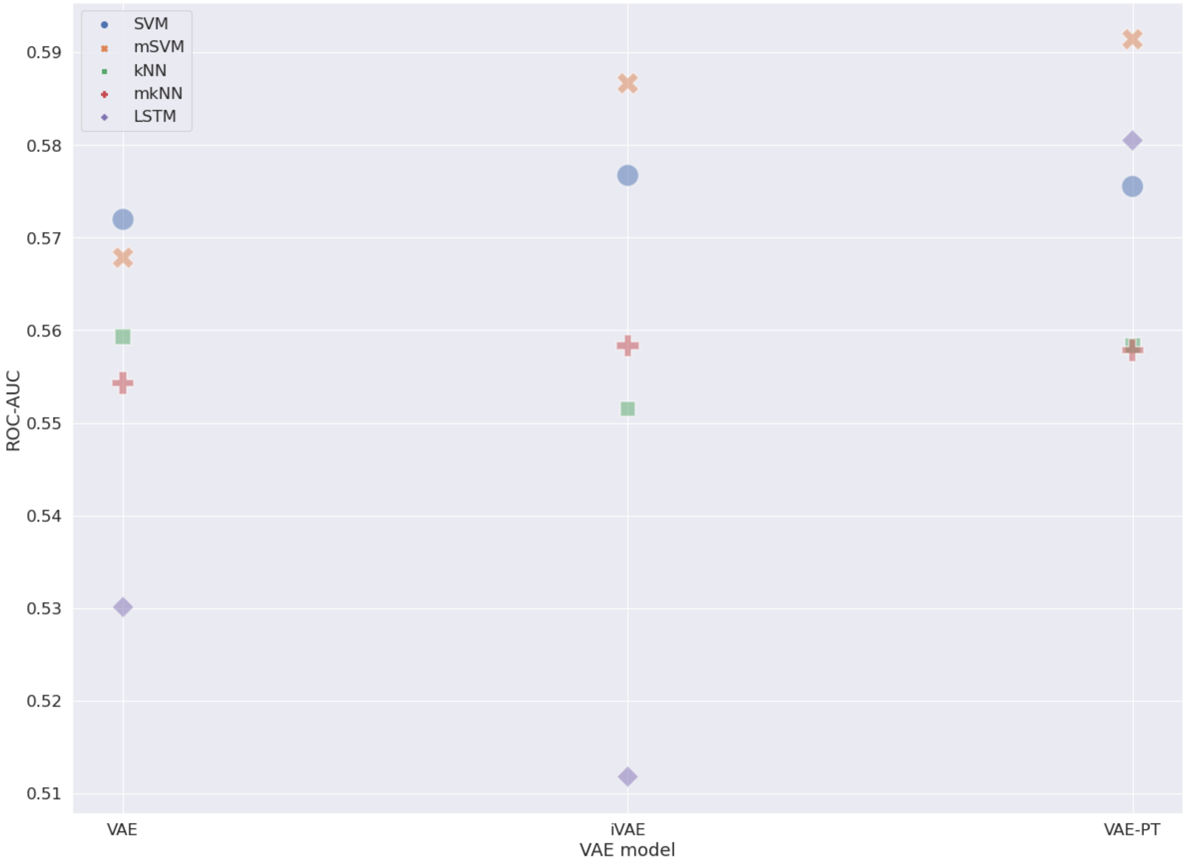


Figure 11: The results for the downstream schizophrenia classification task, the area under the curve for the receiver operating characteristic (ROC-AUC) is shown on the y-axis. The 3 different models are shown on the x-axis: the VAE model, the iVAE model and the pre-trained VAE model (VAE-PT). The pre-training improves the ROC-AUC for the mSVM and LSTM. The iVAE-SVM and iVAE-mSVM outperform all other models, although the results in general are rather disappointing, with a maximum ROC-AUC of 0.592.

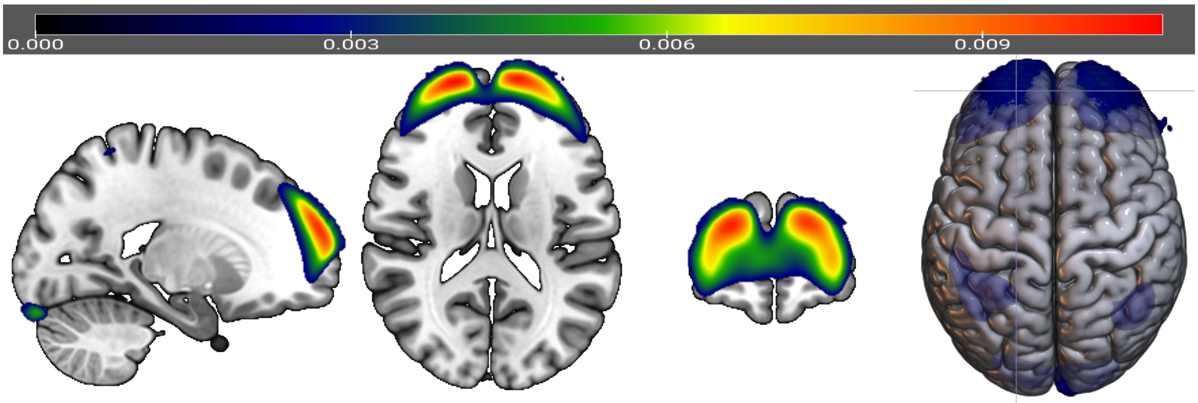


Figure 12: The brain differences in females compared to males, calculated by subtracting the reconstructed average latent representation for males from the reconstructed average latent representation for females. The volume is then thresholded at the highest 80th quantile absolute value. The visualization shows significantly higher activation in the prefrontal cortex and some small increased activation in the occipital lobe.

The VAE generally outperformed the iVAE model, which was likely because the VAE clustered subjects with their timesteps together in the latent space. Including a flexible time prior as part of the objective does not alter the subject-wise clustering, it just results in timesteps being more dispersed around a subject's cluster. Further, the SVM and mSVM outperform all other downstream classification and regression methods, which suggests that the representations allow for linear separation in the latent space for both the sex classification task and the age regression task. The SVM and mSVM perform similarly on tasks which indicates that there is no extra information in the temporal element of the representations, which was also found in similar unpublished work with supervised models [2]. In the unpublished work, a supervised deep learning model trained on the same dataset found that taking the mean over the rs-fMRI time series improved the age regression performance as opposed to using the complete time series. The best performance for the age classification task was achieved by the $512$-dimensional VAE-mSVM with a ROC-AUC of $0.994$, which slightly outperforms the baseline: $0.993$. The best performance for the age regression task is achieved by the $512$-dimensional VAE-SVM with an MAE of $4.014$ years, an R2 score of $0.5288$, and a correlation between the predicted and ground truth ages of $0.727$. This VAE-SVM significantly outperforms the best performing baseline with an MAE of $16.765$ years, an R2 score of $-5.691$, and a correlation between the predicted and ground truth ages of $0.101$. The VAE performs better with larger latent dimensionalities, although the $128$-dimensional VAE did not converge.

The performance on the downstream schizophrenia classification task requires future work. The best ROC-AUC, $0.5914$, was achieved by a pre-trained $256$-dimensional VAE-mSVM. The mSVM outperformed the SVM for all models, which again suggests that the temporal dimension of the representations does not add any valuable information for linear separability. The iVAE and pre-trained VAE perform better than the VAE, which indicates that the flexible time prior and pre-training could be interesting future directions to adapt the VAE towards improved downstream schizophrenia classification.

The results on the age regression and sex classification tasks are promising and the results on the schizophrenia classification task require further work. A common theme for the results seems to be that the temporal dimension of the representations does not add any information that may be valuable for downstream tasks, even when a flexible time prior is introduced. An almost equivalent problem with voxelwise rs-fMRI data also appears in unpublished work for a supervised classification task [2]. Future work should focus on understanding the temporal dynamics of voxelwise rs-fMRI data with respect to downstream tasks. It seems unlikely that the information embedded in the dynamics of the rs-fMRI data would not aid in more meaningful representations. Further, this work finds that subject-wise clustering is a problem that should be tackled in future work to possibly improve all downstream tasks and their pre-training.

# 11. Discussion

The results in this work show that there is great potential for voxelwise rs-fMRI representation learning with a VAE. The representations that are learned by the VAE contain information that allows linear classifiers and regressors to predict the sex and age of a subject with high precision. Further work is required to improve the representations for mental disorders and increase the meaningful information embedded in the dynamics of the representations. The current models seem to perform subject-wise clustering, which may be caused by a small batch size. Other optimizers, learning rate schedules, and larger batch sizes need to be explored further to counter this problem. This also ties into work that needs to be done to make this an efficient solution. An important bottleneck during this project was loading the large data files onto the GPU. Improving data movement and transfer will make experimentation more feasible and may also address the shortcomings of smaller batch sizes. If training times can be reduced, it is also possible to train on larger datasets and perform hyperparameter tuning on a large scale. Larger datasets may also help solve the non-convergence of the VAEs on the downstream schizophrenia classification task. Smaller datasets are more likely to lead to overfitting, especially since voxelwise rs-fMRI data is highly noisy. Not only should future work look at larger datasets, but also move towards models that more efficiently use the information embedded in smaller datasets. Models that incorporate inductive biases and/or forms of regularization are required to move towards meaningful non-linear representations for mental disorders from voxelwise rs-fMRI data. Previous work has already explored the addition of extra regularization in the latent space to improve the disentanglement of the latent space dimensions in VAEs [8, 27].

Although the iVAE did not perform better on the downstream age regression and sex classification tasks, the downstream schizophrenia classification tasks point towards the possibility of future improvement for more complex tasks. The underperformance of the iVAE may be caused by the structure of the latent space and the subject-wise clustering. Future work should explore other auxiliary variables, such as previous timesteps, which were also mentioned in the original work [38]. Auxiliary variables such as previous timesteps may help uncover more meaningful information from the dynamics in the rs-fMRI signal. Another inductive prior that may improve the meaningful information in the temporal dimension of the representations is to parametrize a neural network that is trained in an end-to-end manner and tries to predict future timesteps. This is in line with recent developments in reinforcement learning [23–25]. The work could also try to mimic ICA a little more by imposing a flexible prior based on the dFNC state of the current input volume. Another way to move more towards ICA is to use the time series of an ICA component as a prior for the time series of a certain latent dimension. The number of latent dimensions would have to equal the number of ICA components to obtain a one-to-one mapping, where each latent dimension could be interpreted as the ICA component. Recent work [46] also points to imposing an L1 prior on the state changes between representations and explores a tighter identifiability proof for natural data than Khemakhem et al. [38].

Other future work, certainly for the downstream schizophrenia classification task, could look at improving the classifiers in this work that are used for the downstream classification tasks. The work could also shift towards semi-supervised learning, by adding a classifier and a term in the objective function that optimizes classification accuracy. More powerful or specific classifiers may be able to use the information in the representations and make more accurate predictions. The LSTM used in this work did not outperform any of the linear methods though. Classifiers may also improve on downstream classification tasks when a single representation is learned for a subject. These types of representations could be achieved by training an LSTM to produce a single representation for a subject based on its latent time series. Further, there is work on hierarchical VAEs [45, 65, 74] that may allow for a hierarchy where a subject representation generates the time series representations. These hierarchical VAEs may in general perform better on classification tasks because they can achieve tighter marginal log-likelihood bounds. Flow-based models have also recently been introduced as a method to perform exact inference and train with an exact evaluation of the marginal log-likelihood [6, 13, 14, 41, 45, 66]. Exact evaluation of the marginal log-likelihood and the ability to learn a prior without having to use variational methods can lead to much better generative models. These methods seem to have a high potential for data where a reasonable prior can be constructed, which is true for rs-fMRI data. Flow-based models have also recently been extended to video data, a derivative of which may be useful for future work on voxelwise rs-fRMI representation learning [49].

## 12. Acknowledgements

First and foremost, I want to thank my supervisors for their help and trust in me during this project.I am grateful for the opportunity to work in a field that I am passionate about and I look forward to making meaningful contributions in this field as my research progresses. Thank you to Noah Lewis for giving me tips on how to start my research and for the sparring sessions during my research. Anees Abrol was helpful by sharing his research, which just so happened to be complementary to my own. I am also thankful for having such supportive friends and family. Performing research during a pandemic has sometimes taken a toll, but their support helped keep me motivated.

# Bibliography

[1] Anees Abrol, Zening Fu, Mustafa Salman, Rogers Silva, Yuhui Du, Sergey Plis, and Vince Calhoun. Deep learning encodes robust discriminative neuroimaging representations to outperform standard machine learning. *Nature communications*, 12(1):1–17, 2021.

[2] Anees Abrol, Reihaneh Hassanzadeh, Sergey Plis, and Vince Calhoun. Deep learning in resting-state fmri. *unpublished*, 2021.

[3] CJ Aine, H Jeremy Bockholt, Juan R Bustillo, José M Cañive, Arvind Caprihan, Charles Gasparovic, Faith M Hanlon, Jon M Houck, Rex E Jung, John Lauriello, et al. Multimodal neuroimaging in schizophrenia: description and dissemination. *Neuroinformatics*, 15(4): 343–364, 2017.

[4] Matej Artac, Matjaz Jogan, and Ales Leonardis. Incremental pca for on-line visual learning and recognition. In *Object recognition supported by user interaction for service robots*, volume 3, pages 781–784. IEEE, 2002.

[5] Christian F Beckmann, Clare E Mackay, Nicola Filippini, and Stephen M Smith. Group comparison of resting-state fmri data using multi-subject ica and dual regression. *Neuroimage*, 47(Suppl 1):S148, 2009.

[6] Jens Behrmann, Will Grathwohl, Ricky TQ Chen, David Duvenaud, and Jörn-Henrik Jacobsen. Invertible residual networks. In *International Conference on Machine Learning*, pages 573–582. PMLR, 2019.

[7] B Biswal, F Zerrin Yetkin, Victor M Haughton, and James S Hyde. Functional connectivity Echo-Planar MRI. *Magn Reson Med*, 34(4):537–541, 1995. URL `http://onlinelibrary.wiley.com/doi/10.1002/mrm.1910340409/abstract`.

[8] Christopher P Burgess, Irina Higgins, Arka Pal, Loic Matthey, Nick Watters, Guillaume Desjardins, and Alexander Lerchner. Understanding disentangling in beta-vae. *arXiv preprint arXiv:1804.03599*, 2018.

[9] Vibeke Sorensen Catts, Samantha Jane Fung, Leonora Elizabeth Long, Dispesh Joshi, Ans Vercammen, Katherine Margaret Allen, Stu Gregory Fillman, Loretta Moore, Debora Rothmond, Duncan Sinclair, et al. Rethinking schizophrenia in the context of normal neurodevelopment. *Frontiers in cellular neuroscience*, 7:60, 2013.

[10] Djork-Arné Clevert, Thomas Unterthiner, and Sepp Hochreiter. Fast and accurate deep network learning by exponential linear units (elus). *arXiv preprint arXiv:1511.07289*, 2015.

[11] Bin Dai, Yu Wang, John Aston, Gang Hua, and David Wipf. Connections with robust pca and the role of emergent sparsity in variational autoencoder models. *The Journal of Machine Learning Research*, 19(1):1573–1614, 2018.

[12] Eswar Damaraju, Elena A Allen, Aysenil Belger, Judith M Ford, S McEwen, DH Mathalon, BA Mueller, GD Pearlson, SG Potkin, A Preda, et al. Dynamic functional connectivity analysis reveals transient states of dysconnectivity in schizophrenia. *NeuroImage: Clinical*, 5:298–308, 2014.

[13] Laurent Dinh, David Krueger, and Yoshua Bengio. Nice: Non-linear independent components estimation. *arXiv preprint arXiv:1410.8516*, 2014.

[14] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real nvp. *arXiv preprint arXiv:1605.08803*, 2016.

[15] Yuhui Du, Zening Fu, Jing Sui, Shuang Gao, Ying Xing, Dongdong Lin, Mustafa Salman, Anees Abrol, Md Abdur Rahaman, Jiayu Chen, et al. Neuromark: An automated and adaptive ica based pipeline to identify reproducible fmri markers of brain disorders. *NeuroImage: Clinical*, 28:102375, 2020.

[16] Dumitru Erhan, Aaron Courville, Yoshua Bengio, and Pascal Vincent. Why does unsupervised pre-training help deep learning? In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 201–208. JMLR Workshop and Conference Proceedings, 2010.

[17] Bruce Fischl. Freesurfer. *Neuroimage*, 62(2): 774–781, 2012.

[18] Alex Fornito, Andrew Zalesky, and Michael Breakspear. The connectomics of brain disorders. *Nature Reviews Neuroscience*, 16(3): 159–172, 2015.

[19] KJ Friston, CD Frith, PF Liddle, and RSJ Frackowiak. Functional connectivity: the

principal-component analysis of large (pet) data sets. *Journal of Cerebral Blood Flow & Metabolism*, 13(1):5–14, 1993.

[20] Kunihiko Fukushima. Neocognitron: A hierarchical neural network capable of visual pattern recognition. *Neural networks*, 1(2):119–130, 1988.

[21] Soham Gadgil, Qingyu Zhao, Adolf Pfefferbaum, Edith V Sullivan, Ehsan Adeli, and Kilian M Pohl. Spatio-temporal graph convolution for resting-state fmri analysis. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 528–538. Springer, 2020.

[22] EUGENE V Golanov, SEIJI Yamamoto, and DONALD J Reis. Spontaneous waves of cerebral blood flow associated with a pattern of electrocortical activity. *American Journal of Physiology-Regulatory, Integrative and Comparative Physiology*, 266(1):R204–R214, 1994.

[23] Karol Gregor, George Papamakarios, Frederic Besse, Lars Buesing, and Theophane Weber. Temporal difference variational autoencoder. *arXiv preprint arXiv:1806.03107*, 2018.

[24] David Ha and Jürgen Schmidhuber. World models. *arXiv preprint arXiv:1803.10122*, 2018.

[25] Danijar Hafner, Timothy Lillicrap, Ian Fischer, Ruben Villegas, David Ha, Honglak Lee, and James Davidson. Learning latent dynamics for planning from pixels. In *International Conference on Machine Learning*, pages 2555–2565. PMLR, 2019.

[26] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[27] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. 2016.

[28] Ashley C Hill, Angela R Laird, and Jennifer L Robinson. Gender differences in working memory networks: a brainmap meta-analysis. *Biological psychology*, 102:18–29, 2014.

[29] R Devon Hjelm, Vince D Calhoun, Ruslan Salakhutdinov, Elena A Allen, Tulay Adali, and Sergey M Plis. Restricted boltzmann machines for neuroimaging: an application in identifying intrinsic networks. *NeuroImage*, 96:245–260, 2014.

[30] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[31] Kurt Hornik, Maxwell Stinchcombe, and Halbert White. Multilayer feedforward networks are universal approximators. *Neural networks*, 2(5):359–366, 1989.

[32] Aapo Hyvarinen and Hiroshi Morioka. Unsupervised feature extraction by time-contrastive learning and nonlinear ica. *arXiv preprint arXiv:1605.06336*, 2016.

[33] Aapo Hyvarinen and Hiroshi Morioka. Nonlinear ica of temporally dependent stationary sources. In *Artificial Intelligence and Statistics*, pages 460–469. PMLR, 2017.

[34] Aapo Hyvarinen, Hiroaki Sasaki, and Richard Turner. Nonlinear ica using auxiliary variables and generalized contrastive learning. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 859–868. PMLR, 2019.

[35] Thomas R Insel. Rethinking schizophrenia. *Nature*, 468(7321):187–193, 2010.

[36] Madiha J Jafri, Godfrey D Pearlson, Michael Stevens, and Vince D Calhoun. A method for functional network connectivity among spatially independent resting-state components in schizophrenia. *Neuroimage*, 39(4):1666–1681, 2008.

[37] David B Keator, Theo GM van Erp, Jessica A Turner, Gary H Glover, Bryon A Mueller, Thomas T Liu, James T Voyvodic, Jerod Rasmussen, Vince D Calhoun, Hyo Jong Lee, et al. The function biomedical informatics research network data repository. *Neuroimage*, 124:1074–1079, 2016.

[38] Ilyes Khemakhem, Diederik Kingma, Ricardo Monti, and Aapo Hyvarinen. Variational autoencoders and nonlinear ica: A unifying framework. In *International Conference on Artificial Intelligence and Statistics*, pages 2207–2217. PMLR, 2020.

[39] Jung-Hoon Kim, Yizhen Zhang, Kuan Han, Minkyu Choi, and Zhongming Liu. Representation learning of resting state fmri with variational autoencoder. *bioRxiv*, 2020.

[40] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[41] Diederik P Kingma and Prafulla Dhariwal. Glow: Generative flow with invertible 1x1 convolutions. *arXiv preprint arXiv:1807.03039*, 2018.

[42] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[43] Diederik P Kingma and Max Welling. An introduction to variational autoencoders. *arXiv preprint arXiv:1906.02691*, 2019.

[44] Diederik P Kingma, Danilo J Rezende, Shakir Mohamed, and Max Welling. Semi-supervised learning with deep generative models. *arXiv preprint arXiv:1406.5298*, 2014.

[45] Diederik P Kingma, Tim Salimans, Rafal Jozefowicz, Xi Chen, Ilya Sutskever, and Max Welling. Improving variational inference with inverse autoregressive flow. *arXiv preprint arXiv:1606.04934*, 2016.

[46] David Klindt, Lukas Schott, Yash Sharma, Ivan Ustyuzhaninov, Wieland Brendel, Matthias Bethge, and Dylan Paiton. Towards nonlinear disentanglement in natural data with temporal sparse coding. *arXiv preprint arXiv:2007.10930*, 2020.

[47] Sergey Kolesnikov. Accelerated deep learning rd. `https://github.com/catalyst-team/catalyst`, 2018.

[48] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25:1097–1105, 2012.

[49] Manoj Kumar, Mohammad Babaeizadeh, Dumitru Erhan, Chelsea Finn, Sergey Levine, Laurent Dinh, and Durk Kingma. Videoflow: A flow-based generative model for video. *arXiv preprint arXiv:1903.01434*, 2(5), 2019.

[50] Timothy O Laumann, Abraham Z Snyder, Anish Mitra, Evan M Gordon, Caterina Gratton, Babatunde Adeyemo, Adrian W Gilmore, Steven M Nelson, Jeff J Berg, Deanna J Greene, et al. On the stability of bold fmri correlations. *Cerebral cortex*, 27(10):4719–4732, 2017.

[51] Yann LeCun, Bernhard Boser, John S Denker, Donnie Henderson, Richard E Howard, Wayne Hubbard, and Lawrence D Jackel. Backpropagation applied to handwritten zip code recognition. *Neural computation*, 1(4):541–551, 1989.

[52] Zachary C Lipton, David C Kale, Charles Elkan, and Randall Wetzel. Learning to diagnose with lstm recurrent neural networks. *arXiv preprint arXiv:1511.03677*, 2015.

[53] Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen Awm Van Der Laak, Bram Van Ginneken, and Clara I Sánchez. A survey on deep learning in medical image analysis. *Medical image analysis*, 42:60–88, 2017.

[54] Nikos K Logothetis. The neural basis of the blood–oxygen–level–dependent functional magnetic resonance imaging signal. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 357(1424):1003–1037, 2002.

[55] David JC MacKay and David JC Mac Kay. *Information theory, inference and learning algorithms*. Cambridge university press, 2003.

[56] Usman Mahmood, Md Mahfuzur Rahman, Alex Fedorov, Noah Lewis, Zening Fu, Vince D Calhoun, and Sergey M Plis. Whole milc: generalizing learned dynamics across tasks, datasets, and populations. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 407–417. Springer, 2020.

[57] Dolores Malaspina, Susan Harlap, Shmuel Fennig, Dov Heiman, Daniella Nahon, Dina Feldman, and Ezra S Susser. Advancing paternal age and the risk of schizophrenia. *Archives of general psychiatry*, 58(4):361–367, 2001.

[58] Takashi Matsubara, Koki Kusano, Tetsuo Tashiro, Ken'ya Ukai, and Kuniaki Uehara. Deep generative model of individual variability in fmri images of psychiatric patients. *IEEE Transactions on Biomedical Engineering*, 2020.

[59] Martin J McKeown and Terrence J Sejnowski. Independent component analysis of fmri data: examining the assumptions. *Human brain mapping*, 6(5-6):368–372, 1998.

[60] Karla L Miller, Fidel Alfaro-Almagro, Neal K Bangerter, David L Thomas, Essa Yacoub, Junqian Xu, Andreas J Bartsch, Saad Jbabdi, Stamatios N Sotiropoulos, Jesper LR Andersson, et al. Multimodal population brain imaging in the uk biobank prospective epidemiological study. *Nature neuroscience*, 19(11): 1523–1536, 2016.

[61] Christopher S Monk, Scott J Peltier, Jillian Lee Wiggins, Shih-Jen Weng, Melisa Carrasco, Susan Risi, and Catherine Lord. Abnormalities of intrinsic functional connectivity in autism spectrum disorders. *Neuroimage*, 47(2):764–772, 2009.

[62] Xin Niu, Fengqing Zhang, John Kounios, and Hualou Liang. Improved prediction of brain age using multimodal neuroimaging data. *Human brain mapping*, 41(6):1626–1643, 2020.

[63] National Institute of Mental Health. Schizophrenia, 2020. URL `https://www.nimh.nih.gov/health/topics/schizophrenia/index.shtml`.

[64] Seiji Ogawa, Tso-Ming Lee, Alan R Kay, and David W Tank. Brain magnetic resonance imaging with contrast dependent on blood oxygenation. *proceedings of the National Academy of Sciences*, 87(24):9868–9872, 1990.

[65] Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. *arXiv preprint arXiv:1711.00937*, 2017.

[66] George Papamakarios, Theo Pavlakou, and Iain Murray. Masked autoregressive flow for density estimation. *arXiv preprint arXiv:1705.07057*, 2017.

[67] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019. URL `http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-perform` `pdf`.

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

[69] Han Peng, Weikang Gong, Christian F Beckmann, Andrea Vedaldi, and Stephen M Smith. Accurate brain age prediction with lightweight deep neural networks. *Medical Image Analysis*, 68:101871, 2021.

[70] Fernando Pérez-García, Rachel Sparks, and Sebastien Ourselin. TorchIO: a Python library for efficient loading, preprocessing, augmentation and patch-based sampling of medical images in deep learning. *arXiv:2003.04696 [cs, eess, stat]*, March 2020. URL `http://arxiv.org/abs/2003.04696`. arXiv: 2003.04696.

[71] Sergey M Plis, Md Faijul Amin, Adam Chekroud, Devon Hjelm, Eswar Damaraju, Hyo Jong Lee, Juan R Bustillo, KyungHyun Cho, Godfrey D Pearlson, and Vince D Calhoun. Reading the (functional) writing on the (structural) wall: multimodal fusion of brain structure and function via a deep neural network based translation approach reveals novel impairments in schizophrenia. *Neuroimage*, 181:734–747, 2018.

[72] Ning Qiang, Qinglin Dong, Yifei Sun, Bao Ge, and Tianming Liu. Deep variational autoencoder for modeling functional brain networks and adhd identification. In *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, pages 554–557. IEEE, 2020.

[73] Muhammad Naveed Iqbal Qureshi, Jooyoung Oh, and Boreom Lee. 3d-cnn based discrimination of schizophrenia using resting-state fmri. *Artificial intelligence in medicine*, 98:10–17, 2019.

[74] Ali Razavi, Aaron van den Oord, and Oriol Vinyals. Generating diverse high-

fidelity images with vq-vae-2. *arXiv preprint arXiv:1906.00446*, 2019.

[75] Michal Rolinek, Dominik Zietlow, and Georg Martius. Variational autoencoders pursue pca directions (by accident). In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12406–12415, 2019.

[76] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3): 211–252, 2015.

[77] Tim Salimans and Diederik P Kingma. Weight normalization: A simple reparameterization to accelerate training of deep neural networks. *arXiv preprint arXiv:1602.07868*, 2016.

[78] Hui Shen, Lubin Wang, Yadong Liu, and Dewen Hu. Discriminative analysis of resting-state functional connectivity patterns of schizophrenia using low dimensional embedding of fmri. *Neuroimage*, 49(4):3110–3121, 2010.

[79] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1): 1929–1958, 2014.

[80] Jing Sui, Rongtao Jiang, Juan Bustillo, and Vince Calhoun. Neuroimaging-based individualized prediction of cognition and behavior for mental disorders and health: methods and promises. *Biological psychiatry*, 2020.

[81] Carol A Tamminga, Elena I Ivleva, Matcheri S Keshavan, Godfrey D Pearlson, Brett A Clementz, Bradley Witte, David W Morris, Jeffrey Bishop, Gunvant K Thaker, and John A Sweeney. Clinical phenotypes of psychosis in the bipolar-schizophrenia network on intermediate phenotypes (b-snip). *American Journal of psychiatry*, 170(11):1263–1274, 2013.

[82] RAPIDS Development Team. *RAPIDS: Collection of Libraries for End to End GPU Data Science*, 2018. URL `https://rapids.ai`.

[83] Michael Tschannen, Olivier Bachem, and Mario Lucic. Recent advances in autoencoder-based representation learning. *arXiv preprint arXiv:1812.05069*, 2018.

[84] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

[85] Stefan Van Der Walt, S Chris Colbert, and Gael Varoquaux. The numpy array: a structure for efficient numerical computation. *Computing in science & engineering*, 13(2): 22–30, 2011.

[86] Svyatoslav Vergun, Alok Deshpande, Timothy B Meier, Jie Song, Dana L Tudorascu, Veena A Nair, Vikas Singh, Bharat B Biswal, Mary Elizabeth Meyerand, Rasmus M Birn, et al. Characterizing functional connectivity differences in aging adults using machine learning on resting state fmri data. *Frontiers in computational neuroscience*, 7:38, 2013.

[87] Elizabeth Reisinger Walker, Robin E McGee, and Benjamin G Druss. Mortality in mental disorders and global disease burden implications: a systematic review and meta-analysis. *JAMA psychiatry*, 72(4):334–341, 2015.

[88] Xiaodi Zhang, Eric Maltbie, and Shella Keilholz. Spatiotemporal trajectories in resting-state fmri revealed by convolutional variational autoencoder. *bioRxiv*, 2021.

[89] Qingyu Zhao, Nicolas Honnorat, Ehsan Adeli, Adolf Pfefferbaum, Edith V Sullivan, and Kilian M Pohl. Variational autoencoder with truncated mixture of gaussians for functional connectivity analysis. In *International Conference on Information Processing in Medical Imaging*, pages 867–879. Springer, 2019.