FRI Data Science Project Competition 2024
Final report

Classifying Psychiatric Disorders

Ahmet Çalış, Manfred Gonzalez-Hernandez and Joaquín Figueira

Abstract

In recent times the need for a neurobiological basis to treat psychosis spectrum disorders has become apparent, as the clinical treatment of these disorders still relies on self-reported behavioral analysis of the patients which has proven to be inaccurate. In this work we use data from fMRI scans to predict the disorders of a set of 435 psychosis spectrum disorder patients, with the ultimate goal of finding a small latent space of the fMRI data in which the different disorders and their symptoms are clearly distinguishable. To achieve this goal we combine and apply different approaches from traditional Machine Learning and Network Analysis, including some dimensionality deduction techniques. The results suggest that it's possible to accurately differentiate a patient from the healthy control group and map this distinction in a meaningful latent space, but predicting specific disorders (schizophrenia, schizoaffective disorder and bipolar disorder with psychosis) within the patient group and symptomatic scales proved unsuccessful with the data and methodology applied.

Keywords
Data Science, Psychiatry, fMRI, B-SNIP, GBC, FC, Network Analysis, Machine Learning, PANSS

Advisors: Prof. Dr. Jure Demšar

Introduction

Given the high levels of complexity of the human brain and its associated afflictions, diagnosing and treating psychiatric disorders is a notoriously challenging endeavor. This is specially true in the psychosis spectrum of disorders (PSD), where the literature shows that, due to its greater symptomatic variability, there is a pressing need for more complex and individualized patient care.

One of the biggest impediments to this sort of treatment is the fact that traditional clinical approaches used to diagnose and treat these disorders relay on behavioral and self reported observations of the patients symptoms, such as the Positive and Negative Syndrome Scale (PANSS). Although a successful neural mapping across canonical schizofrenia (SZP) symptoms has been found in previous works [1], the field still lacks an accurate neural mapping for the full spectrum of pyschosis disorders. Finding such a mapping could be used to provide more accurate forms of treatments and diagnosis with a sound biomolecular basis.

In this context, this project aims to bridge this behavioral and neurological gap by developing Machine Learning (ML) approaches to relate different patient fMRI scans with their corresponding diagnosis, using data obtained by the Bipolar and Schizophrenia Network for Intermediate Phenotypes consortium (B-SNIP) [2]. The data consists of a set of PANSS scores, diagnosis and resting state fMRIs scans from patients diagnosed with schizophrenia (SZP), schizoaffective disorder (SADP) and bipolar disorder with psychosis (BPP), and a healthy control group (CON). The ultimate objective of this work is to build a classifier that can accurately predict the specific disorder of a patient using their corresponding fMRI data, and use it to find a small latent space representation of the neural data that highly correlates to PSD symptoms.

Methods

Data preprocessing and dimensionality reduction

An fMRI scan records the brain activity of a patient through blood oxygenation measurements. In these measurements the fMRI scanner intrinsically divides the brain into different regions (formally called voxels) according to a pre-defined spatial resolution, which tipycally results in a subdivision of the brain in around 90,000 voxels. The result is a time series of 3-D images of blood oxygenation levels for each voxel sampled with an approximate time resolution of 1.5 seconds during 20-30 minutes, which all combined amounts to huge quantities of data for even a single patient.

As the high dimensionality of this data representation would probably lead to high degrees of overfitting and excessively high runtimes, we used several dimensionality reduction techniques to make it more digestible. First, we merged voxels into regions of interest (ROIs), for which we used two parcellations: the Glasser parcellation which has 718 ROIs and the Cole-Anticevic parcellations whith 12 ROIs.

With this first reduction, we obtained a time series of images with a more compact spatial resolutions. However, the data volume was still intractable due to the size of the time dimension of the series. To tackle this we used two other techniques for time dimensionality reduction: Functional Connectome (FC) and Global Brain Connectivity (GBC). The first approach, FC, measures the total correlation between each pair of ROIs across time. In other words, for a parcellation with r ROIs, the data is reduced to an $\mathbb{R}^{r\times r}$ matrix. The second approach,

GBC, computes the average correlation between one ROI and all the others across time, resulting in a r dimensional vector.

Traditional Machine Learning on GBC data

We'll introduce this section with some nomenclature. There are two specific tasks we wanted to solve in this work. The first task was to classify whether a person had a disorder or not, we'll refer to this task as **health status classification**. The second task is to identify the specific disorder of a person, if any. We'll refer to this task as **disorder classification**. To tackle them our first approach was to use three learners to build one model each for both tasks. Additionally, as we wanted to experiment with the two different parcellations explained in the previous section, we built specific models for each of them. This amounted to a total of 12 classifiers.

For the learning algorithms we chose the following: random forests (RF), XGBoost and multilayer perceptrons (MLP). As random forests are known to perform relatively well with little parameter and feature extraction procedures, we decided to use them as a general baseline. Furthermore, since we're working in the medical domain, explainability is key, which is why we decided to use XGBoost as our main classification model, as it provides both decent performance and explainability. Additionally, it can perform well in low sample environments. To test more complex feature interactions, we decided to use Multi Layer Perceptrons (MLP).

To build the models and evaluation we used a train-test split of the data with 20% test size. We performed hyper-paramer tunning using the train split and for evaluation we computed the models' accuracy in both splits. The results obtained with the previously described models can be seen in Table 1. Note that we omit the results for the Cole-Anticevic parcellation as they were significantly lower even for health status classification (less than 0.7 accuracy). We can see that all models achieve very high accuracies for health status classification, even the baseline RF model. However, for disorder classification we weren't able to improve the accuracy compared with the baseline.

Metric	Task	XGBoost	RF	MLP
Accuracy (Tr)	Status	1.0	1.0	1.0
Accuracy (Ts)	Status	0.984	0.977	0.98 ± 0.02
Accuracy (Tr)	Disorder	1.0	1.0	1.0
Accuracy (Ts)	Disorder	0	0.586	0.55 ± 0.08

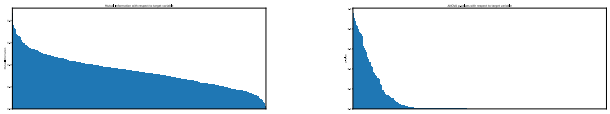
Table 1. Classification results before feature selection. In the table *Tr* stands for training and *Ts* stands for test. Furthermore, the **task** column indicates the task: *Status* refers to health status classification and *Disorder* to disorder classification.

Feature Selection

A major issue we found with our methodology was a major imbalance in the data between the 718 features and the insufficient number of samples (638). This was our motivation to create a more involved feature selection process. Furthermore, in implementing the previous approach we noted that performance dropped significantly with increased data compression. Consequently, we chose to work with 718 ROI Glasser parcellation data.

First, we checked for constant and quasi constant values. Next, we performed feature correlations and removed 84 features using a correlation threshold of 0.8. Additionally, we used mutual information, ANOVA test and univariate model performances.

As seen in Figure 1a above, there was no clear cut off point for mutual information, so we removed features with mutual information



(a) (b)

Figure 1. Mutual information and ANOVA test results.

below 0.5. Figure 1b shows results for ANOVA, it revealed more features that had no relationship with the target value. We removed 129 features with help of these two processes.

For the univariate model performance based feature selection, we trained decision trees for each feature and checked whether this features helped the model to perform better than a random baseline. This process removed 7 features at the end.

All of the techniques we utilized so far helped us to remove 222 features in total. We decided to continue with step forward feature selection (SFS) from this point. All previous steps are very helpful for SFS since its computationaly expensive process. SFS identified 5 key features at the end. These five features are labeled in the dataset as: X534, X484, X426, X284, and X684. We used these 5 features to check model performances.

Using this reduce set of features we retrained our XGBoost models using the same methodology as before. The results can be observed in Table 2, where we can see that in the disorder classification task feature selection didn't affect performance on the test set (it produced a drop in accuracy of only 0.01) and also helped to reduce overfitting, as performance in the training set decreased to a more realistic 0.73. Furthermore, health status classification results remained unchanged as well.

Task	Accuracy (Tr)	Accuracy (Ts)
Status	0.988	0.96 ± 0.01
Disorder	0.73	0.55 ± 0.04

Table 2. Classification results after feature selection using XGBoost. The notation is the same as in Table 1.

In conclusion, both disorder and health status classification results indicate that only 5 features are sufficient to achieve the same level of performance as using all of them. We hypothesize that most features are uninformative due to excessive data compression during preprocessing. Furthermore, in terms of absolute performance, we can see that all models achieve very high accuracies for health status classification, even the baseline RF model, but we weren't able to improve the performance in the disorder classification task.

Network Analysis on FC data

In the upcoming stage of our research, we develop a graph-based framework to enhance feature extraction methods and find approaches for multi-class classification accuracy. We employed an undirected graph for each patient from a functional connectivity matrix by adding edges based on a threshold determined by the standard deviation of the absolute values in the matrix. Initially, the function calculates the threshold as a specified multiple (given by the standard deviation (std) multiplier) of the standard deviation of the absolute connectivity values. Each node, corresponding to an entity in the matrix, is added to the graph. Then, for every pair of nodes, an edge is created if the absolute value of the connectivity strength between them exceeds the calculated threshold, excluding self-loops by ensuring the nodes

are different. The weight of each edge corresponds to the original connectivity strength. This method results in a graph where only the strongest connections, those significantly above the average connectivity strength, are represented, highlighting the most prominent functional relationships within the network.

By doing this we explored the features of two types of graphs, one graph with a higher standard deviation multiplier where the amount of edges between nodes will be considerably lower than the second graph that had a smaller standard deviation multiplier. This can be seen in tables 3 and 4. Features like the average degree measures the average number of connections each node (brain region) has, it helps in understanding how interconnected the brain regions are. Then features like the clustering coefficient provides insight into the local connectivity patterns and the tendency of brain regions to form tightly connected groups. The average path length feature could simulate the actual efficiency of information transfer within the brain. In brain networks, a large connected component could suggest that a significant portion of the brain regions are functionally integrated, which is essential for coordinated brain function. The initial idea is that these features could describe each patient's network and apply machine learning models over these features. But the final performance of such models was poor, even worse than the random baseline.

Via these graphs we wanted to analyze the sub-graphs orbits using a different architecture. In this case we linked the nodes with the top K connectivity values of each of the regions, to be able to see how these sub-graphs expressed different information of the brain. To do so, we experimented with the arithmetic agreement similarity using the orbit counts in Orca [3]. After applying the so called agreement similarity over all these graphs we built the figure 2 sorting the axes by the different groups or labels that each of the patient belongs to. The initial idea is that patients with the same label will highlight higher similarity near to the diagonal of the figure, but this did not happen.

Table 3. Summary of graph statistics for graphs with $STD = 2$. Notation: m = Number of Edges, $\langle K \rangle$ = Average Degree, $\langle C \rangle$ = Average Clustering, $\langle PL \rangle$ = Average Path Length, CC = Size of Largest CC.

Statistic	BPP	CON	SADP	SCZP
m	82137	78250	78832	84185
$\langle K \rangle$	228.79	217.97	219.59	234.50
$\langle C \rangle$	0.55	0.55	0.54	0.56
$\langle PL \rangle$	1.69	1.71	1.70	1.68
CC	717.83	717.88	717.84	717.89

Table 4. Summary of graph statistics for graph with $STD = 4.5$ Same notation as in table 3.

label	BPP	CON	SADP	SCZP
m	5083.51	1672.45	2622.66	8815.35
$\langle K \rangle$	14.16	4.66	7.31	24.56
$\langle C \rangle$	0.11	0.09	0.09	0.12
$\langle PL \rangle$	4.31	4.00	4.11	4.36
CC	146.41	118.44	147.94	156.89

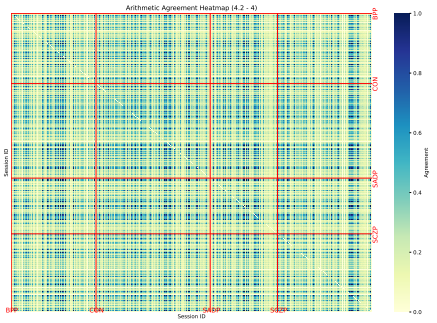


Figure 2. Each pixel is the comparison of one session id FC graph with it's Arithmetic Agreement based on Orca orbits of 5 nodes. Each red line makes a boudary between session ids of different groups

Latent Space representation

As stated in the introduction of this paper, our ultimate goal was to find a mapping between PSD symptoms (encapsulated in the PANSS scores of the patients) and their neurological expression. To this end we developed a latent space representation technique to visually separate and cluster the different groups of patients. Ideally, a successful implementation of this approach should be able to correctly separate the different diagnosis into distinct clusters in a low dimensional space, allowing us to see possible similarities between them based on the distance and shapes of their clusters.

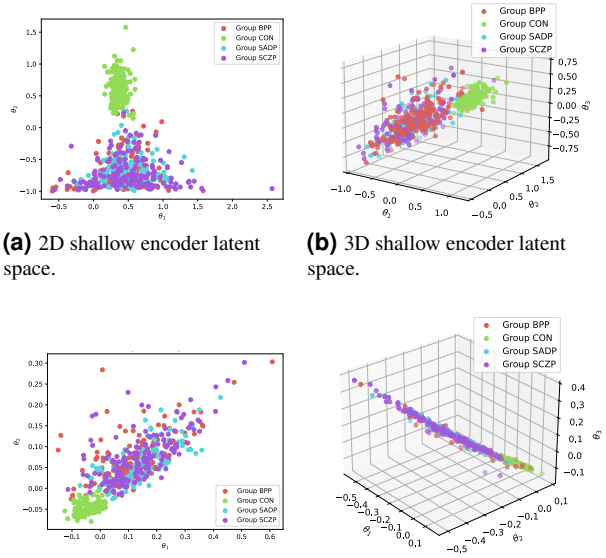
The technique we developed uses the 718 ROI parcellation GBC data to predict the PANSS scores of the patients using a Deep Neural Network (DNN) architecture with a central embedding layer. This layer has a small number of outputs that represent precisely the latent space of the GBC data. In a sense, we wanted to replicate an autoencoder architecture in which the input and output are (theoretically) correlated, but they're clearly distinct.

The main hypothesis behind this implementation is that, if the diagnosis and associated symptoms are truly characterized by distinct brain patterns, then the network, while optimizing to reproduce the PANSS score, should be able to construct a latent or internal representation where disorders are clustered nicely.

We tried several configurations of this base architecture using different number of hidden layers and different sizes of the embedding layer. In particular, we experimented with 2D and 3D embedding layers and one and two hidden layers in the encoder and decoder sections of the network. As we wanted to replicate the PANSS scores as precisely as possible we used Means Squared Error as loss function for the model.

In Figure 3 we can visualize the results for the four architecture configurations we implemented. In it we can see the embeddings of the GBC data with 718 ROI labeled with their respective diagnosis. In all of the shallow representations we can see clear clusters separating patients from the control group. However, it seems the deep encoders compress the data too excessively by combining all the features multiple times. This results in very flat/uni-dimensional latent representations. Furthermore, none of the variations we tried were able to find a suitable representation where all the different groups are clearly differentiable.

As this results proved unsuccessful in segregating the different patient groups, we applied an additional dimensionality reduction technique known as t-SNE to see if we could improve them. In figure

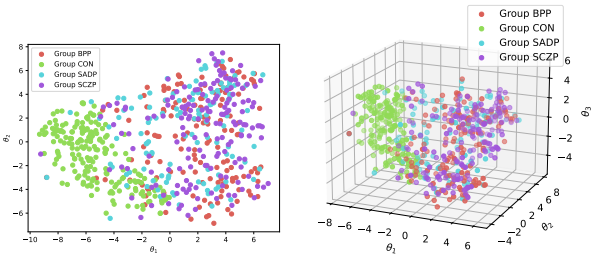


(a) 2D shallow encoder latent space. **(b)** 3D shallow encoder latent space.

(c) 2D deep encoder latent space. **(d)** 3D deep encoder latent space.

Figure 3. Latent space representations of the Glasser parcellation GBC data using each of the encoder architectures implemented. The datapoints are labeled with their respective disorder group.

4 we can see the results, where we can observe even less clearly defined clusters (although there does seem to be an ordering distinction between patients and control groups). This serves as counterfactual evidence of the fact that there does seem to be correlations between behavioral and neural data which are helping the DNN cluster the groups more efficiently.



(a) 2D latent space. **(b)** 3D latent space.

Figure 4. Latent space representation of the Glasser parcellation GBC data using t-SNE with 2 and 3 output dimensions. The datapoints are labeled with their respective disorder group.

Discussion

The results we obtained suggest that the problem of identifying the binary health status of a patient using biomarker data is solvable, as already suggested in the literature. We've proven this in multiple ways by building direct classifiers on the GBC data with high levels of precision and by sucessfuly encoding the data in a 2D and 3D latent space where control and patient groups are clearly discernible. We've also identified a relatively small number of regions of the Glasser parcelation where most of the differentiation between patient and healthy control group seems to originate.

In relation to the Network Analysis of the FC data, we've determined that the specific combination of data and extensive analytical techniques we used is unable to tackle the problems of health status and disorder classification. We hypothesize that the main problem with our approach is the features extraction procedure, as extracting features manually from the graphs is challenging and it may require more advanced approaches such as graph embeddings or graph neural networks that are outside of the scope of this work.

In respects to disorder classification, all the approaches we experimented with proved unsuccessful. We believe that the main reason for this is the large amounts of compression applied to the data. Several of our results seem to support this conclusion: **1)** the reduction in performance produced by the Cole-Anticevic Parcelation suggests that choosing a correct parcellation is very important to achieve accurate results; **2)** the low amount of ROIs (only 0.6% of all ROIs) that we identified were most correlated with predictive performance seem to suggest that GBC is too aggressive a method to maintain relevant information from all the ROIs.

Another reason for the low accuracy in this task may be that the diagnosis themselves are simply not reflected in the imaging data, a conclusion supported by the literature [2].

Future Work

Our results suggest that to solve the problem of disorder classification of PSD new dimensionality reduction and representation techniques for fMRI data need to be developed. On the one hand, it'd be beneficial to develop different spatial reduction techniques and neural atlases with a focus on maintaining as much as possible the spatial variability of the data. We believe more fine grained atlases will be specially useful in this respect, as advances in machine learning in recent years make it feasible to train more complex algorithms such as DNN on this high dimensional data.

Furthermore, we believe that, to solve the task of disorder classification, temporal dimensionality reduction techniques such as GBC should be avoided. This is because such general approaches probably discard too much information in the compression of the feature interactions through time. Instead, we believe it'd be more effective to use sequence transformers on the time series data after applying only spatial dimensionality reduction through parcellation. We believe this approach can more meaningfully encode complex feature interactions in the data through time and hence generate better results.

Finally, using more advanced network analysis techniques on the Functional Connectome data could provide better insights into the neurological study of PSD. In particular, it could help in finding improved graph representations of the data that could facilitate the use of ML techniques to find highly correlated ROIs in the brain that can be targeted by specific medications.

References

[1] Ji Chen, Kaustubh R. Patil, and et al. Neurobiological divergence of the positive and negative schizophrenia subtypes identified on a new factor structure of psychopathology using non-negative factorization: An international machine learning study. *Biological Psychiatry*, 87(3):282–293, February 2020.

[2] Brett A. Clementz, John A. Sweeney, Jordan P. Hamm, Elena I. Ivleva, Lauren E. Ethridge, Godfrey D. Pearlson, Matcheri S. Keshavan, and Carol A. Tamminga. Identification of distinct psychosis biotypes using brain-based biomarkers. *American Journal of Psychiatry*, 173(4):373–384, April 2016.

[3] Tomaž Hočevar and Janez Demšar. A combinatorial approach to graphlet counting. *Bioinformatics*, 30(4):559–565, 12 2014.