Classification and Prediction from Multimodal Neuroimaging Data in the Context of Schizophrenia Treatment

Bc. Jakub Svoboda

Supervisor: Mgr. Jaroslav Hlinka, PhD.
Consultant: Mgr. Barbora Rehák Bučková, PhD.

MSc. Programme: Medical Electronics and Bioinformatics
Branch of Study: Signal Processing

Faculty of Electrical Engineering
Czech Technical University in Prague

May 2024

Thesis Supervisor:

Ing. Mgr. Jaroslav Hlinka, PhD.
Department of Complex Systems
Institute of Computer Science of the Czech Academy of Sciences
Pod Vodárenskou věží 271
182 00 Prague 8
Czech Republic

Thesis Consultant:

Mgr. Barbora Rehák Bučková, PhD.
Donders Centre for Cognitive Neuroimaging
Trigon building
Kapittelweg 29
6525 EN Nijmegen

MASTER‘S THESIS ASSIGNMENT

I. Personal and study details

Student's name: **Svoboda Jakub** Personal ID number: **483542**

Faculty / Institute: **Faculty of Electrical Engineering**

Department / Institute: **Department of Circuit Theory**

Study program: **Medical Electronics and Bioinformatics**

Specialisation: **Signal processing**

II. Master’s thesis details

Master’s thesis title in English:

Classification and Prediction from Multimodal Neuroimaging Data in the Context of Schizophrenia Treatment

Master’s thesis title in Czech:

Klasifikace a predikce z multimodálních neurozobrazovacích dat v kontextu léčby schizofrenie

Guidelines:

Schizophrenia is a chronic, severe and profoundly disabling disorder. For every 100 individuals with schizophrenia, only 1 or 2 individuals per year meet the recovery criteria, and approximately 14% are expected to recover over 10 years, with poor functional outcome for 27% of patients [1]. There is an urgent need to develop predictive models of outcome to be applied in the initial stages of illness and thus optimize and intensify intervention programs to avoid an aversive outcome. Functional outcomes are difficult to predict solely on the basis of the clinical features, but Magnetic Resonance Imaging (MRI) holds promise for improved stratification of patients [2]. The ultimate aim is to develop tools to predict the functional outcome of schizophrenia from multimodal neuroimaging, clinical and cognitive measurements taken early after disease onset. To overcome limitations due to high dimensionality of data, one should combine robust machine-learning tools, data-driven feature selection and theory-based brain network priors [3,4]. It is key to investigate the potential of the modalities themselves, in order to find out to what extent the (expensive) increase of sample size itself would help the ML performance, in particular whether the maximum potential of the features has been practically reached and more extensive feature engineering is needed.

1. Conduct a literature review on the detection of schizophrenia and prediction of therapeutic outcome from unimodal and multimodal neuroimaging data.
2. Investigate dependence of the separating plane between the groups on their theoretical distributions and sample size. Evaluate on data the separating potential in individual modalities and across modalities.
3. Prepare an appropriate structured dataset from data obtained as part of the ESO study conducted at the National Institute of Mental Health.
4. Design a suitable fusion pipeline of available multimodal neuroimaging data for machine learning.
5. Evaluate the possibilities of classification and regression tasks in schizophrenia on the multimodal neuroimaging dataset.
6. Validate results on an independent dataset.

Bibliography / sources:

[1] Menezes, N. M., Arenovich, T., & Zipursky, R. B. (2006). A systematic review of longitudinal outcome studies of first-episode psychosis. Psychol Med, 36(10), DOI: https://doi.org/10.1017/S0033291706007951
[2] McGuire, P., & Dazzan, P. (2017). Does neuroimaging have a role in predicting outcomes in psychosis? World Psychiatry, 16(2), doi: 10.1002/wps.20426
[3] Rehák Bučková, B., Mareš, J., Škoch, A., Kopal, J., Tintěra, J., Dineen, R., Řasová, K., & Hlinka, J. (2022). Multimodal-neuroimaging machine-learning analysis of motor disability in multiple sclerosis. Brain Imaging and Behavior. https://doi.org/10.1007/s11682-022-00737-3
[4] Calhoun VD, Sui J. Multimodal fusion of brain imaging data: A key to finding the missing link(s) in complex mental illness. Biol Psychiatry Cogn Neurosci Neuroimaging. 2016 May;1(3):230-244. doi: 10.1016/j.bpsc.2015.12.005. PMID: 27347565; PMCID: PMC4917230.

Name and workplace of master's thesis supervisor:

Ing. Mgr. Jaroslav Hlinka, Ph.D. Institute of Computer Science, The Czech Academy of Sciences, Prague

Name and workplace of second master's thesis supervisor or consultant:

Mgr. Barbora Rehák Bučková Donders Centre for Cognitive Neuroimaging (en), Nijmegen, The Netherlands

Date of master's thesis assignment: **20.09.2023** Deadline for master's thesis submission: **24.05.2024**

Assignment valid until: **16.02.2025**

Ing. Mgr. Jaroslav Hlinka, Ph.D.	doc. Ing. Radoslav Bortel, Ph.D.	prof. Mgr. Petr Páta, Ph.D.
Supervisor's signature	Head of department's signature	Dean's signature

III. Assignment receipt

The student acknowledges that the master's thesis is an individual work. The student must produce his thesis without the assistance of others, with the exception of provided consultations. Within the master's thesis, the author must state the names of consultants and include a list of references.

.	
Date of assignment receipt	Student's signature

Declaration

I declare that the presented work was developed independently and that I have listed all sources of information used within it in accordance with the methodical instructions for observing the ethical principles in the preparation of university theses.

In Prague, May 2024

.....

Bc. Jakub Svoboda

Abstract

With the paradigm shift to multimodal machine learning, hope arises for an improvement in the prediction of functional outcome of schizophrenia and other disorders. As magnetic resonance imaging (MRI) can produce several types of views into the human brain, it is suitable for the development of multimodal pipelines. The Early-Stage Schizophrenia Outcome study aims to collect multimodal, longitudinal neuroimaging and clinical data of first-episode schizophrenia patients and healthy controls in order to construct a predictive framework of functional outcome and treatment trajectory. From the data collected within this project up until now, we have assembled two datasets ("IKEM" dataset: 161 subjects, "NÚDZ" dataset: 204 subjects) of features derived from several brain MRI modalities (structural, functional, and diffusion). We have surveyed state-of-the-art multimodal machine learning literature and conducted exploratory analyses our data, assessing the baseline predictive potential of primarily linear models in detecting schizophrenia and predicting clinical features. Functional MRI features achieved the highest predictive potential in schizophrenia detection, and also show the highest rate of accuracy improvement with increasing sample size. Conversely, regression on scores of the positive and negative symptoms has not shown applicable results as of yet, regardless of modality. When integrating multiple modalities, we found the highest predictive performance in a nonlinear predictor stacked from several unimodal logistic classifiers. With the findings from our analyses, we emphasize the need to understand the potential and limitations of both the data and methods in designing individualized predictive systems.

Keywords: machine learning, exploratory analysis, magnetic resonance imaging, multimodal integration, schizophrenia

Abstrakt

Se změnou paradigmatu k multimodálnímu strojovému učení přichází naděje na zlepšení kvality predikce prognózy u schizofrenie a dalších poruch. Magnetická rezonance (MRI) je schopna zachytit různé typy obrazu mozku, je tak vhodná pro vývoj multimodálních predikčních metod. Studie Early-Stage Schizophrenia Outcome si klade za cíl nashromáždit multimodální, longitudinální neurozobrazovací data pacientů s první epizodou schizofrenie a zdravé kontrolní skupiny, a sestavit rámec pro predikci funkčního výsledku a trajektorie léčby. Z dosud naměřených dat jsme sestavili dvě datové sady ("IKEM" dataset: 161 subjektů, "NÚDZ" dataset: 204 subjektů) sestávající z příznaků z několika modalit mozkových obrazů (strukturální, funkční a difuzní). Provedli jsme rešerši současné literatury o multimodálním strojovém učení a exploratorní analýzy prediktivního potenciálu primárně lineárních modelů při detekci schizofrenie a predikci klinických ukazatelů. Příznaky z funkční MRI vykazují nejvyšší prediktivní potenciál při detekci schizofrenie. Naopak regresní predikce hodnot škály pozitivních a negativních symptomů dosud neukázala využitelné výsledky. Při integraci více modalit dosáhl nejvyššího prediktivního výkonu nelineární prediktor složený z několika unimodálních logistických klasifikátorů. S poznatky z našich analýz zdůrazňujeme potřebu pochopit potenciál a limitace dat, modalit a metod při návrhu individualizovaných prediktivních systémů.

Klíčová slova: strojové učení, exploratorní analýza, magnetická rezonance, multimoální integrace, schizofrenie

Acknowledgements

I want to express my gratitude to the many people that helped me directly or indirectly pursue my master's degree and goals in general. Firstly, I want to thank the super-supervisor duo Dr. Jaroslav "Cyborg" Hlinka and Dr. Barbora "Baša" Rehák Bučková for supervising me and trying to direct my chaotic energy towards actually producing some results instead of jumping from one idea to another. Both Jarda and Baša have provided me with enormous insight and I suspect my now very strong scepticism is mostly their doing. Baša is also responsible for digging up the data from the depths of the NÚDZ data stores. In the process of my work, I've also had the pleasure to discover the Institute of Computer Science at the Czech Academy of Sciences and all of the lovely people there, with whom I gladly procrastinated over coffee. I also owe a big thanks to my family, for supporting me and making sure I did not perish. Finally, I want to thank all of my wonderful friends, who listened to me and encouraged me.

Bc. Jakub Svoboda

List of Figures

List of Tables

Acronyms

AAL Automated Anatomical Labeling atlas

AD axial diffusivity

AI artificial intelligence

ALFF amplitude of low-frequency fluctuations

BP bipolar patient

CV cross-validation

DT duration of treatment

DUP duration of untreated psychosis

dwMRI diffusion-weighted magnetic resonance imaging

EEG electroencephalography

ESO Early-Stage Schizophrenia Outcome

FA fractional anistropy

FC functional connectivity

FEP first-episode psychosis

FES first-episode schizophrenia

fMRI functional magnetic resonance imaging

GAF Global Assessment of Functioning

HC healthy control

HYDRA Hyper Database of Recognizable Associations

IKEM Institute of Clinical and Experimental Medicine

JHU John Hopkins University atlas

M.I.N.I. Mini-International Neuropsychiatric Interview

MD mean diffusivity

MRI magnetic resonance imaging

NUDZ National Institute of Mental Health

PANSS Positive and Negative Symptom Scale

PCA principal component analysis

RD radial diffusivity

ROI region of interest

SBM surface-based morphometry

SC structural connectivity

sMRI structural magnetic resonance imaging

SVM support vector machines

SZ schizophrenia patient

VBM voxel-based morphometry

WHOQOL World Health Organization Quality of Life

Contents

Preface

Machines rising

The recent developments in artificial intelligence (AI) and medical technology make the case for computer-assisted healthcare and precision medicine ever more hopeful. Optimizing hospital processes, automatically detecting abnormal patterns from medical devices, or enabling telemedicine for remote monitoring and advice are just some examples of smart technology already helping save and improve lives. An important area of research with much advancement potential is the prediction of diagnosis, treatment trajectory, and outcome.

Eluding illness

Mental illnesses still await a significant breakthrough in this regard, as they are much less tangible than conditions with a primarily physiological or neurologic nature. *Schizophrenia*, a mental condition that can severely warp one's perception of reality from the norm, is one such illness that would greatly benefit from computer-assisted treatment. However, it has no clear etiology, and all of its symptoms overlap with those of other psychiatric disorders. It is up to trained specialists to determine the diagnosis, which can be quite subjective, as even professionals with the same training can perceive the patient's symptoms and behavior differently. As such, it is desirable to find some objective biomarkers of the disease, which could help ascertain the diagnosis or even recommend individualized treatment. The hope is that early intervention and precisely targeted treatment can greatly improve the prognosis of the patient.

Capture and predict

To build predictive frameworks, we use features or biomarkers derived from data. In an attempt to investigate the objective markers for schizophrenia in the brain, researchers turn to neuroimaging. Brain imaging techniques emerged along with the wave of novel

medical imaging technology in the 20th century and enabled us to capture and map the structure and function of the brain. To name a few, electroencephalography (EEG) measures the electrical activity of the brain cortex through the scalp, and magnetic resonance imaging (MRI) uses strong electromagnetic fields in various configurations to measure the tissue as a structural image, time series of neuronal metabolic activity, and so on. These methods use different physical principles for measurement and so capture slightly or wildly different types of information about the studied tissue. When building models and predictors from neural data, we rely on these empirically known properties to exploit them well and improve predictive and interpretative power.

Enter multimodality

The term *modality* denotes a set of features that contain specific information about the measured entity. Multimodality represents a general paradigm shift happening in machine learning fields based on the belief that working with multiple data types and integrating different views of studied objects can lead to improved prediction. Examples of such approach include the DALL-E models integrating images with text to allow generating imagery from textual prompts (Ramesh et al. 2022), or ImageBind, which creates shared representation for a number of modalities – images, text, audio, depth, thermal, and inertial measurement unit (e.g. gyroscopic) data (Girdhar et al. 2023).

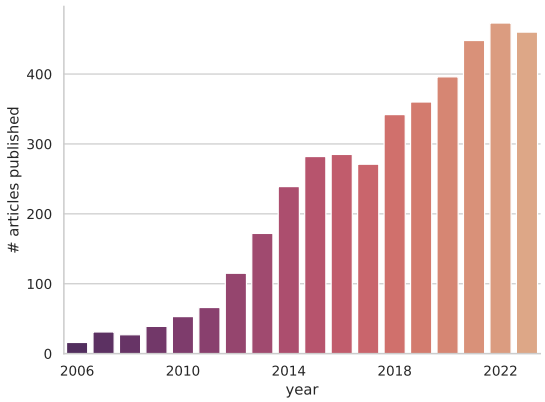

Figure 1: Results of a search in the Scopus database with the terms *multimodal neuroimaging* (author, abstract, and keyword fields) showing yearly numbers of article publications in this area.

Neuroimaging also provides a variety of ways to inspect the brain, the neuroimaging community started adopting the multimodal paradigm, and so in the last decade there has been a surge in research focused on multimodal fusion (Azam et al. 2022). Thus more studies have begun measuring multiple MRI modalities during a single magnetic resonance imaging recording session, and there are also efforts to concurrently record func-

tional magnetic resonance imaging (fMRI) and EEG, exploiting both the spatial resolution of fMRI and temporal resolution of EEG (Jorge, Van Der Zwaag, and Figueiredo 2014; Piorecka et al. 2022). Methods and pipelines integrating multiple imaging modalities are gathered under the umbrella term *multimodal fusion*.

Multimodal confusion

While the idea of "the more modalities, the better" does seem intuitive, multimodality comes with its own class of challenges. Firstly, there comes the MRI economy and practicality (Ooi et al. 2024). Scanning sessions are quite expensive at the baseline; adding recording blocks for additional modalities increases this price tag. More importantly, subjects have to spend more time in the scanner, which is not overly comfortable and, in fact, may be distressing for some. This can cut the sessions short (Quirk et al. 1989), or render the recordings unusable due to strong motion artifacts. In the already data-starved biomedical field, multimodal datasets usually count even fewer recordings than unimodal datasets. While there are possibilities of aggregating datasets from multiple sites, the problem lies in differing scanners, acquisition protocols, and demographics on site, and harmonizing these factors adds another layer of complexity. Additionally, with more modalities come more features, resulting in larger dimensionality. Models with a large number of parameters tend to overfit to small training sets, resulting in poorer predictions on unseen data (Berisha et al. 2021). This means greedy approaches such as neural networks or blindly using all features in prediction are not preferable. So, more complex feature sets might potentially include novel information, as well as novel noise, to derail our algorithms. As such, we have to reduce, simplify, and regularize our models, to recover as much signal as possible while constrained by the sample size available.

ESO up the sleeve

The Early-Stage Schizophrenia Outcome (ESO) study is an ongoing effort to construct a longitudinal multimodal dataset for schizophrenia. The main aim of the project is to optimize care by predicting diagnosis, guiding treatment, and tracking illness trajectory to improve the quality of life and functional outcomes of patients. Acquiring a multitude of MRI modalities, as well as collecting many clinical variables, the project intends to construct a complex clinical profile of the volunteering subjects. The data acquisition is actively taking place formerly at the Institute of Clinical and Experimental Medicine (IKEM) and currently at the National Institute of Mental Health (NUDZ) in the Czech Republic and aims to form one the most comprehensive collections of longitudinal multimodal data from schizophrenia patients and healthy controls. The objectives of the study

are defined as follows:

Objective 1: Robust multimodal feature extraction To establish an effective software, hardware and data framework to store and analyze large multimodal data.

Objective 2: Unimodal data analyses To develop and apply methods for efficient prediction of probable treatment outcome from individual data types.

Objective 3: Multimodal data analyses To improve accuracy of trajectory prediction by combining data from multiple modalities.

Objective 4: Clinical Application To establish simplified disease models providing interpretable prediction of the disease progression trajectory.

The task before us

The goal of this thesis is to perform a multimodal analysis of a subset of the ESO dataset, positioning us between Objectives 2 and 3. With the difficulties and limitations outlined above in mind, we intend to study model behavior with respect to dimensionality, intermodality, and sample size. We resort to using simple linear models to inspect the dataset; constructing a complex and intricate model that would predict a diagnosis with the best accuracy possible is out of the scope of this thesis.

The official tasks of this thesis are defined as follows:

Task 1 Conduct a literature review on the detection of schizophrenia and prediction of therapeutic outcome from unimodal and multimodal neuroimaging data.

Task 2 Investigate dependence of the separating plane between the groups on their theoretical distributions and sample size. Evaluate on data the separating potential in individual modalities and across modalities.

Task 3 Prepare an appropriate structured dataset from data obtained as part of the ESO study conducted at the National Institute of Mental Health.

Task 4 Design a suitable fusion pipeline of available multimodal neuroimaging data for machine learning.

Task 5 Evaluate the possibilities of classification and regression tasks in schizophrenia on the multimodal neuroimaging dataset.

Task 6 Validate results on an independent dataset.

Structure of this work

The tasks defined above will be fulfilled in the chapter structure, which is as follows:

Research & Review (Task 1) First, we review fundamental concepts related to this thesis. Then, we survey the academic literature related to machine learning on multimodal MRI data and see what we can expect in terms of the performance of machine learning models and the limitations of the multimodal approach.

Prepare Dataset (Task 3) From the ESO project database at NUDZ, we collect and filter out relevant data to derive feature sets that will comprise our multimodal dataset.

Explore & Design (Task 2, 4, and 5) Exploring the assembled dataset, we think of ways how to properly engineer and combine different feature sets for the machine learning tasks.

Hypothesize and Validate (Task 6) Using the models discovered previously, we lay out models and hypotheses that we test on an independent dataset.

Ponder & Conclude We evaluate attained results and discuss them in the context of the reviewed research, concluding the thesis.

Chapter 1

Research & Review

1.1 The split mind

Schizophrenia is a mental disorder that warps one's perception of reality. It can severely impair the patient's functioning in everyday life and negatively affect their place in society. The definition of schizophrenia today is still not clear cut, and the development of the concept up to this point was a rocky path throughout history. One term largely used throughout the ages was lunacy, and it encompassed a multitude of today's differentiated mood and psychotic disorders, including schizophrenia. Important milestones on this journey include French psychologist Benedict Morel defining the term dementia praecox (meaning "premature dementia"); the work of Emil Kraepelin, who defined a disease with similar symptoms to today's schizophrenia under the same term; Eugen Bleuer following up on this work and introducing his own term, schizophrenia, at the beginning of the twentieth century (Johnstone 2003). In the same century, the first standardized diagnostic criteria were established by major health organizations: the Diagnostic and Statistical Manual of Mental Disorders by the American Psychiatric Association (now in its 5th edition) and the International Classification of Diseases by the World Health Organisation (now in the 11th version).

The global prevalence of schizophrenia has been recently estimated to be roughly 289 cases per 100,000 people in 2019 (Solmi et al. 2023). Signs of schizophrenia include positive symptoms: psychosis (e.g., hallucinations) and delusions (e.g., paranoia, grandiosity), and negative symptoms: blunted affect, social withdrawal, catatonic behavior (abnormal movement patterns), and so on (Schultz, North, and Shields 2007). The symptoms are diverse, yet none are specific to the illness. In addition, schizophrenia patients are still stigmatized, and their human rights are often violated, which may worsen the condition (Kelly 2005).

Pharmacological treatment of schizophrenia goes back to the 1950s with the first gen-

eration of antipsychotics (also referred to as neuroleptics) (Schultz, North, and Shields 2007). The first antipsychotic developed was chlorpromazine, and to this day, antipsychotic dosages are often expressed as chlorpromazine equivalents in milligrams (Patel et al. 2013). First-generation (typical) antipsychotics have proven to be effective, at least in reducing the positive symptoms, though they do carry a range of adverse side effects with them. The characteristic adverse effects of typicals are extrapyramidal side effects, which comprise various motor dysfunctions (Arana 2000). While second-generation (atypical) antipsychotics have come shortly after the first, the distinction between typicals and atypicals has been contested, as the atypical drugs do not seem to improve on efficacy or tolerability significantly. Rather, the distinction lies in different side effects, with atypicals alleviating extrapyramidal effects and having more pronounced metabolic effects (Leucht et al. 2009). This is to say, there is no silver bullet regarding drug prescription, and it is not uncommon for patients to switch medication multiple times during their treatment to find a good trade-off between efficacy and tolerability for their case (Bitter et al. 2008).

1.2 Magnetic resonance imaging – see inside

The unique property of MRI is that it enables noninvasive acquisition of three-dimensional reconstruction of internal tissue. It exploits the behavior of hydrogen atoms in a strong magnetic field and the abundance of water in the human body. The atoms are aligned by the strong magnet. Then, coils disrupt this configuration with radiofrequency waves, and fluctuations created by the atoms returning to their aligned state are measured as signal. Depending on the tissue the water is bound to, multiple signal intensities arise, creating an image (Brown and Semelka 2011). A number of MRI sequences exist for perturbing the field, leading to different imaging contrasts or modalities. In this thesis, we will refer only to specific neuroimaging contrasts – structural, functional, and diffusion-weighted imaging.

1.2.1 Structural MRI

The fundamental modality is structural magnetic resonance imaging (sMRI), capturing the anatomy of brain tissue. Three important contrasts that show up in sMRI are white matter, gray matter, and cerebrospinal fluid (showing as black, i.e. low signal). These tissues can be analyzed with methods such as voxel-based morphometry (VBM) or surface-based morphometry (SBM) (Ashburner and Karl J. Friston 2000; Goto et al. 2022), allowing us to compute and compare measures like gray matter volume or cortical thickness. Structural imaging can reveal injuries, tumors, anatomical abnormalities, atrophy of tissue, and others. Technically, sMRI is key for any MRI study, regardless of the primary modality of interest, as structural images serve as a reference in the preprocessing steps of images

from all modalities (Manjón 2017; Esteban et al. 2019). To minimize findings due to individual differences such as skull volume, standardized brain templates have been devised by averaging a large number of structural images. When images are acquired for a study, they are registered to this template.

1.2.2 Functional MRI

Brain activity can be measured by fMRI. It cannot capture the underlying electromagnetic activity of the neurons, but there is a proxy measure – neuronal metabolism. As active neurons require more oxygen, a hemodynamic response delivers more oxygenated blood into their immediate surroundings, changing magnetic properties in the corresponding voxel and thus showing a different intensity. Functional imaging's strength lies in its greater spatial resolution compared to EEG. EEG records the electrical activity of the brain projected to the scalp, whereas fMRI can capture metabolic activity anywhere inside the brain down to a few millimeters. The downside is that the temporal resolution is not very impressive (which is the strength of EEG). The temporal resolution, usually around two seconds per volume, is not only limited by the scanner's capabilities but also by the hemodynamic response itself since the duration from stimulus onset to peak response is in order of seconds Figure 1.1.

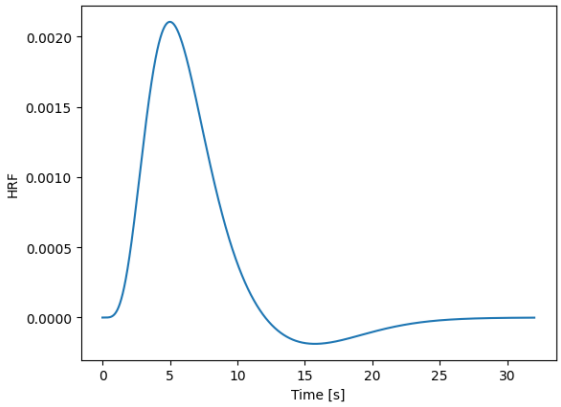

Figure 1.1: A sketch of the hemodynamic response function (HRF) model, as implemented in the statistical parametric mapping package (*SPM12 Software - Statistical Parametric Mapping* 2024). The HRF peaks around five seconds from onset.

1.2.3 Diffusion-weighted MRI

Diffusion-weighted imaging displays the complex structure of axonal highways; in other words, the structure of white matter in the brain. As axons are responsible for neuronal communication, diffusion-weighted magnetic resonance imaging (dwMRI) serves as a means of inspecting structural connectivity and white matter integrity, which can be sensitive to pathological changes. The principle of the method is based on the motion of water

molecules in the brain. The constraint of motion of the molecules depend on the cell composition in a certain place (represented here by a voxel). In some voxels, the molecules can move freely in most directions (anisotropic motion), but in and around neuronal axons, the motion is dominantly constrained to be along the axon. So, the imaging protocol works by applying magnetic gradient of several strengths in several directions, acquiring images in a number of directions. From these images, the diffusion of water molecules in each voxel can be measured (Hagmann et al. 2006). The methods of estimating the white matter tracts or fibers are termed tractography. With the tracts estimated, we can then compute structural connectivity between regions of interest, by tracing and counting fibers connecting those regions (Jeurissen et al. 2019; Bach et al. 2014). Another class of features obtainable from dwMRI images are tensor measures (diffusion tensor imaging). In each voxel, a tensor is computed which represents an ellipsoid that describes the motion possibilities at that voxel. This tensor has eigenvalues associated with it, that are then used to compute the diffusion tensor measures. For instance, fractional anisotropy measures the relative degree of water diffusion anisotropy in a voxel, with a minimum value of 0 representing spherical diffusion and a maximum value of 1 representing linear diffusion (Le Bihan et al. 2001).

1.3 Atlases of the (brain) world

Throughout the history of investigating the brain, different parcellations have been devised in order to organize its structure anatomically or functionally. MRI atlases represent such divisions as voxel masks - 3D images where the voxel value indicates whether the respective voxel belongs to a respective region/structure in the atlas (Cabezas et al. 2011). The region labels can be probabilistic, expressing numerically the likelihood of belonging to a region, or deterministic, giving a hard indication. Here we describe two specific atlases, which are also used in our project - the Automated Anatomical Labeling atlas (AAL) and the John Hopkins University atlas (JHU).

1.3.1 AAL gray matter atlas

The AAL is an atlas and a software package that labels the gray matter regions of the cortex. Originally, a T1 volume of a single subject was parcellated into gray matter 90 regions (45 for each hemisphere), and 26 subcortical regions, totalling 116 labeled regions (N. Tzourio-Mazoyer et al. 2002). In the subsequent versions, AAL2 (Rolls, Joliot, and Nathalie Tzourio-Mazoyer 2015) and AAL3 (Rolls, C.-C. Huang, et al. 2020), tweaking some regions and including new ones, make the total number of regions 120 and 166, respectively. Our study utilizes only the 90 gray matter regions from AAL1 (Table A.1).

1.3.2 JHU white matter atlas

The ICBM-DTI-81 white matter atlas has been created at the John Hopkins University by averaging DTI scans from 81 subjects and manually delineating 50 anatomical structures (Table A.2) known from histological studies of white matter, as documented by Oishi et al. (2008).

1.4 Clinical examinations

In order to rate the condition or symptom severity of a patient in a quantifiable and standard way, numerous clinical scales and instruments have been devised, rating disorder-specific symptoms or general indications. Clinical scales are used to assess the initial state and illness trajectory of a patient in clinical practice, research, and clinical trials of treatments. Here, we provide short descriptions of four such scales, which were recorded as a part of the ESO project. For the purposes of this thesis, we only deal with the Positive and Negative Symptom Scale (PANSS) scale.

1.4.1 MINI

The Mini-International Neuropsychiatric Interview (M.I.N.I.) is a short, approximately 15-minute structured interview developed to be compatible with diagnostic criteria for psychiatric disorders in the ICD-10 and DSM-IV manuals. It aims to be brief while maintaining accuracy. It is designed to capture routine information so non-specialist personnel can conduct the interview. It is intended both for the clinical and research setting (Sheehan et al. 1998).

1.4.2 BREF

The World Health Organization Quality of Life (WHOQOL) BREF questionnaire is the abbreviated version of WHOQOL-100. It is useful for brief clinical assessments while being highly correlated with the more extensive questionnaire. It surveys four domains of quality of life: physical health, psychological, social relationships, and environment. The examinee is asked to answer each question on a scale of 1 to 5 (worst to best). The answers are then plugged into four equations (one for each domain), computing the final scores (WHO 2004).

1.4.3 GAF

The Global Assessment of Functioning (GAF) scale comprises Axis V in the multiaxial assessment system in the *DSM-IV* (1994). It is a 100-point scale divided into deciles, each

interval representing a symptomatic and functional component. The examiner works with the interval, which reflects the worse of the two components; e.g., if the symptom severity falls between 70 and 80 but functioning falls between 60 and 70, the examiner will select a score they deem appropriate in the range 60-70. GAF is not diagnosis-specific and covers the range from a severely disabled (1) to a healthy (100) individual. Although there are some points of controversy, such as higher sensitivity to the subjectivity of the interviewer, it has been adopted in several local health systems and is a common instrument in outcome studies (Aas 2010).

1.4.4 PANSS

PANSS (Kay, Fiszbein, and Opler 1987) is one of the oldest yet still one of the most widely used scales for assessing the severity of schizophrenia. A trained examiner conducts an approximately 50-minute interview with the patient, studies reports from family and care personnel and compiles this information into the rating. The scale rates thirty symptoms in total, split into three groups: positive symptoms, negative symptoms, and general psychopathology (enumerated in Table 1.1). The positive and negative scales include 7 symptoms each, and the general psychopathology scale includes 16. The examiner rates each symptom on a scale of 1 to 7 in terms of intensity. The minimum score the examinee can attain is 30, the maximum is 210 (the lower, the better).

Table 1.1: PANSS subscales and symptoms

Scale	Symptoms
Positive symptoms	Delusions, Conceptual disorganization, Hallucinatory behavior, Excitement, Grandiosity, Suspiciousness/persecution, Hostility
Negative symptoms	Blunted affect, Emotional withdrawal, Poor rapport, Passive/apathetic social withdrawal, Difficulty in abstract thinking, Lack of spontaneity and flow of conversation, Stereotyped thinking
General psychopathology	Somatic concern, Anxiety, Guilt feelings, Tension, Mannerisms and posturing, Depression, Motor retardation, Unusual thought content, Disorientation, Poor impulse control, Preoccupation, Poor attention, Active social avoidance, Disturbance of volition, Impaired judgment and insight, Uncooperativeness

1.5 Statistical methods

Note that we use notation and formulas for population statistics throughout the text when convenient, however, in practice, we use sample statistics.

1.5.1 Standardization

In order to have comparable statistics across features, as well as to ascertain good behavior of some algorithms, we *standardize* features to zero mean and unit variance. The *standard score* (z-score) of a data point expresses in terms of standard deviations how the raw data point deviates from the mean of the overall data. It is defined as

$$Z = \frac{X - \mu}{\sigma}. \tag{1.1}$$

1.5.2 Testing hypotheses

Statistical tests allow us to asses significance of our findings when analysing data. The *null hypothesis* (H_0) often marks the condition that there is no difference or effect present, such as no linear relationship between two variables. The *alternative hypothesis* (H_1) states the opposite. The output of a test is some statistic and a *p*-value. The *p*-value tells us the probability of observing the found or even more extreme value of the statistic, given that the null hypothesis is true. To make a decision if a finding is statistically significant, we set a *significance level* α, with the conventional level being $\alpha = .05$, or lower. In our analyses, we will use this conventional significance level.

1.5.3 Errors, power, and corrections

When conducting a test, the significance level α has an important meaning: it is the level of acceptable false positive probability. False positive rate is also known as the *Type I error*: *detecting* a significant result when it is *not* there. False negativity or *Type II error*, the probability of which is denoted by β, means *failing* to detect a significant result when it *is* there. On the other hand, the ability of a test to *detect* a significant result when it *is* there, is called *power*. It is defined as the complement of Type II error probability:

$$power = 1 - \beta. \tag{1.2}$$

In the case where multiple tests are conducted in attempt to answer a single question, the error compounds for every test conducted. This compound error is known as the *family-wise error rate* (FWER). Under independence of the hypotheses, the FWER is given by:

$$FWER = 1 - (1 - \alpha)^m, \tag{1.3}$$

where m denotes the total number of tests conducted. Assuming α, for $m = 2$ the FWER will be 0.0975, for $m = 3$ it will be 0.1426. It becomes apparent that even with a small number of tests, the compound error becomes unacceptable. A number of *corrections* were devised to alleviate this issue. The most well-known and very widely used is also one of the most conservative ones – the *Bonferroni* correction. This procedure simply divides the original significance level for a single test by the total number of tests conducted to answer the research question

$$\alpha_m = \frac{\alpha}{m}. \tag{1.4}$$

Alternatively, we can multiply the *p*-values by m, keeping α fixed, to achieve the same effect. While Bonferroni indeed diminishes the probability of false detections (in fact, controls the FWER at values below the prescribed α), it also does so for true detections, meaning it has lower power. Other corrections have a varying tradeoff between FWER control and power, and it is up to the researcher to choose and justify the use of a respective procedure. Although corrective procedures can serve well to control for false findings, a notion often omitted is that they ought not be used blindly, to needlessly hinder power. For instance, when doing exploratory analyses or post-hoc tests, correcting or overly conservative correcting in this scenario could be detrimental to the idea of the task. As such, Bonferroni correction (and even the *p*-value per se) has been questioned in the academic sphere (Wasserstein and Lazar 2016). Our approach is the following: we will use uncorrected significance levels when exploring data or performing post-hocs; however, when we want to make a confident statement when reporting statistical questions, we will use the Bonferroni correction.

1.5.4 Correlations

Statistical relationships between pairs of variables observed together can be measured by correlation coefficients.

Pearson's (product-moment) correlation coefficient measures the degree of linear relationship between variables. For random variables X and Y, Pearson's ρ is defined as their covariance normalized by the product of their standard deviations

$$\rho_{XY} = \frac{\mathrm{cov}(X,\ Y)}{\sigma_X \sigma_Y}. \tag{1.5}$$

Spearman's (rank) correlation coefficient measures the degree of monotonicity of the relationship. The formula is equivalent to Pearson's correlation, except that the values of the variables are first converted into their rank, and then plugged into the formula

$$\rho_{XY} = \frac{\mathrm{cov}(\mathrm{rank}(X),\ \mathrm{rank}(Y))}{\sigma_{\mathrm{rank}(X)} \sigma_{\mathrm{rank}(Y)}}. \tag{1.6}$$

1.6 Machine learning models

In the field of machine learning, models are built on assumptions and trained on available data. They can then predict e.g. the properties of unseen data or discover hidden associations in a dataset. In contrast to traditional group analysis in statistics, which aims to find a difference between populations as a whole, predictive ML models can perform inference on an individual level. In prediction, a model learns a relationship between predictors and a target variable. If this variable is numerical, the task is called regression; if it is categorical, the task is then called classification.

1.6.1 Linear regression

A fundamental tool in any analysis, linear regression models aim to find coefficients to model a linear relationship between a target variable and its predictors. It does so by finding coefficients of the linear combination of the predictors

$$Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p + \varepsilon, \tag{1.7}$$

where Y is the target variable, X_i the predictors, β_i the coefficients, and the residuals ε express the difference between the predicted and true value of Y. The interpretation of the coefficients is that β_i represents the average change in Y for a unit change in X_i, leaving all other predictors fixed.

There are now many methods to obtain in some sense optimal estimates $\hat{\beta}$ of the coefficients β of a linear model, but the classical one is called *ordinary least squares*, which minimizes the sum of squares of the differences between data points and their predicted values. This criterion is expressed as

$$\mathcal{L} = \sum_{i=1}^{n} ||y_i - \hat{\beta}_0 - \sum_{j=1}^{m} \hat{\beta}_j x_{ij}||^2. \tag{1.8}$$

1.6.2 Logistic regression

In a binary classification scenario, with classes labeled 0 and 1, the probability of a data point belonging to class 1 is p_1, and the probablitity of it being class 0 is $p_0 = 1 - p_1$. The fraction of probabilities is called *odds*, and by taking the (natural) logarithm of odds we obtain *log-odds*

$$\frac{p_1}{p_0} = \frac{p_1}{1 - p_1}. \tag{1.9}$$

Logistic regression is a linear model *in parameters*: it models the log-odds as a linear combination of the predictors

$$\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p = \ln\left(\frac{p_1}{1-p_1}\right). \tag{1.10}$$

The model expresses the probability of class 1 via the sigmoid (logistic) function; solving eq. (1.10) for p_1, we obtain the formula

$$p(x) = p_1(x) = \frac{1}{1-\exp^{-(\beta_0+\beta_1 x_1+\ldots+\beta_p x_p)}}. \tag{1.11}$$

One of the ways to estimate model parameters are via maximum likelihood estimation.

1.6.3 Principal component analysis

principal component analysis (PCA) aims to find the directions of the largest variance of the data. These directions are represented by a vector basis, where the vectors are ordered by the amount of variance they capture. This basis is found by eigendecomposition of the covariance matrix of the data. An *eigenvector* v of matrix A fulfills the following condition

$$Av = \lambda v, \tag{1.12}$$

where λ is the corresponding *eigenvalue*. The procedure is often preceded by taking the data matrix X and standardizing its features in columns (assuming the examples are stored in rows), as PCA is sensitive to feature scaling. Then, the covariance matrix is constructed

$$\Sigma = X^T X, \tag{1.13}$$

and its eigenvectors V are found

$$\Sigma = V \Lambda V^T, \tag{1.14}$$

where Λ denotes a diagonal matrix with eigenvalues on the diagonals corresponding to eigenvectors in the columns of V. Plotting the variance captured by each component, we often see that the first few principal components capture much of the useful information and thus decide to keep only a small number of the total number of components; this can be seen both as a form of compression and filtering.

1.6.4 Bias and variance

The integral part of a model is its assumptions, which influence its flexibility. The degree of flexibility relates to the model's ability to adapt to data and plays a central role in the

bias-variance tradeoff. Stronger assumptions can give the model a lot of predictive and interpretative power, but a gross violation of these assumptions can be detrimental to performance. Such models are known to be inflexible or biased, to have a high bias (not to be confused with the bias term in linear models). In contrast, models with generally fewer assumptions, many parameters, and high adaptability are flexible; they have a high variance. The training score of flexible models tends to be higher, as they are better molded to the training set. This is a double-edged sword, though, because these estimators can easily overfit to the details of the training set, leading to insufficient generalization and performance in the real world (referring to new, unseen data). The challenge is then to find a classifier with flexibility fitting the nature of the data at hand. The bias-variance tradeoff is expressed by the formula decomposing total error into the bias and variance components, as well as irreducible error (Geman, Bienenstock, and Doursat 1992):

$$\varepsilon_{total} = \varepsilon_{bias} + \varepsilon_{variance} + \varepsilon_{noise} \tag{1.15}$$

This relation is visualized in Figure 1.2. Sample size plays a big role in this problem: it is more acceptable to choose highly flexible estimators such as convolutional neural networks if there are many examples, like in the MNIST handwritten digit database (Lecun et al. 1998). Though some overparametrized neural network architectures seem to empirically challenge the bias-variance paradigm (Neal et al. 2019), that seems to be the case when the ratio of sample size to model capacity is large and thus avoids the neuroimaging case. As such, less flexible models such as linear models are abundantly used in imaging studies, especially with the typically small samples.

1.6.5 Dimensionality

With large numbers of features, the *curse of dimensionality* starts to be a problem, which leads to troubles in convergence, overfitting, or sampling issues. Reducing the number of dimensions is one of the ways of dealing with the curse of dimensionality. Feature selection is a family of methods aiming to select a subset of variables most valuable for prediction. Some examples would be forward/backward step-wise selection, where in each step, a variable is added into the model if it improves a certain criterion in comparison with other variables; very popular are shrinkage methods, such as ridge or lasso, which incorporate shrinking of the model coefficients directly in the training phase, leading to penalization of less important variables. The lasso specifically zeroes the unimportant variables. Another feature engineering approach is decomposition, with PCA being the standard among linear techniques. It is important to keep in mind that the number of principal components does not have to be the same as the original number of features. The number of components is upper bound by either the number of features or the sample size, whichever is smaller. This

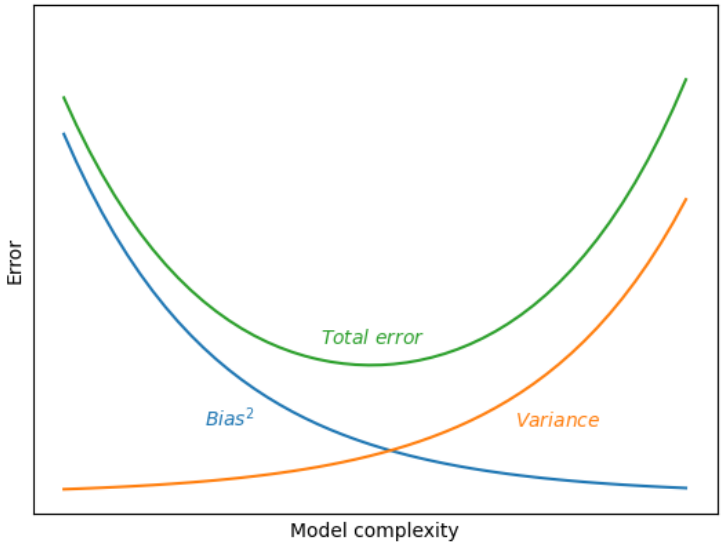

Figure 1.2: Complex models have less bias (rigidity) and have more variance (flexibility). As such, with more complexity, the error due to bias diminishes, and the error due to variance grows. In order to minimize the total error, the aim is to find balanced levels of bias and variance in the model.

ties another layer of complexity to PCA applied to small high-dimensional datasets because every additional training example increases the total number of components obtained by PCA (and also improves the overall estimate of covariance). Nonlinear methods include widespread methods such as kernel PCA, multi-dimensional scaling, uniform manifold approximation and projection, or t-stochastic neighbor embedding. These methods, while more flexible, are more suited to specific conditions and scenarios, e.g., data being spread on a specific, high-dimensional nonlinear manifold.

1.6.6 Regularization

Both simple and complex models benefit from regularization, which aims to prevent overfitting. There are various methods to achieve this, but the most well-known are the ridge and lasso methods, based on penalizing large (absolute) values of model coefficients. This can be done in models using any sort of gradient optimization to find the coefficients. Gradient descent optimizes a loss function based on the optimization criterion for the model, and the penalty is implemented by the addition of a coefficient penalty term. The term is a sum of all the coefficients with a norm applied to them. For the ridge, that is the L2 norm; for the lasso, it is the L1 norm. For least squares linear regression, the extended loss looks

as follows:

$$\mathcal{L}_{ridge} = \sum_{i=1}^{n} ||y_i - \beta_0 - \sum_{j=1}^{m} \beta_j x_{ij}||^2 + \sum_{j=0}^{m} \|\beta_j\|_2^2 \tag{1.16}$$

$$\mathcal{L}_{lasso} = \sum_{i=1}^{n} ||y_i - \beta_0 - \sum_{j=1}^{m} \beta_j x_{ij}||^2 + \sum_{j=0}^{m} \|\beta_j\|_1 \tag{1.17}$$

where $\beta_0, ..., \beta_m$ are the model coefficients, and

$$||x||_p = (|x_1|^p + |x_2|^p + \ldots + |x_n|^p)^{\frac{1}{p}} \tag{1.18}$$

denotes the Minkowski vector p-norm. These methods are called shrinkage methods, as they shrink the size of the coefficients. Lasso specifically tends to zero out some coefficients, effectively performing feature selection. Another popular method, called dropout, has been introduced in neural networks. This method randomly nullifies some neurons (coefficients) in each iteration, forcing other neurons to evolve more robustly in that iteration of the training.

1.6.7 Orthogonalization

There are a number of ways of dealing with confounding variables to minimize their influence on analysis. A simple but data-hungry approach would be confounder stratification. Subjects are analyzed separately in strata or bins of the confounder where the variable values are relatively similar. That is not practical for our dataset size. Another approach is called confounder orthogonalization (also known as Frisch-Waugh-Lovell or regression decomposition theorem). There, for each dependent or independent variable of interest, a linear regression model is constructed where said variable is used as a response and the confounding variables as predictors. The residuals of the model are used as new values for the variable. This diminishes the effect of the expected linear relationship with the confounders, as it is subtracted from the useful variable. Suppose we wanted to orthogonalize age and sex from the variable X. The model would be represented by the equation

$$X = \beta_0 + \beta_{age} c_{age} + \beta_{female} c_{female} + \varepsilon. \tag{1.19}$$

where c stands for a confounder variable and $c_{male}, c_{female} \in \{0, 1\}$ specifically are dummy variables standing in for the categorical variable of c_{sex}. The procedure is visualized in Figure 1.3. This model is a composite of two lines with the same slope and different biases. In the case that the age range was wider and age was more heterogeneous between the sexes, an interaction term could be introduced into the equation, providing the model with two different slopes as well.

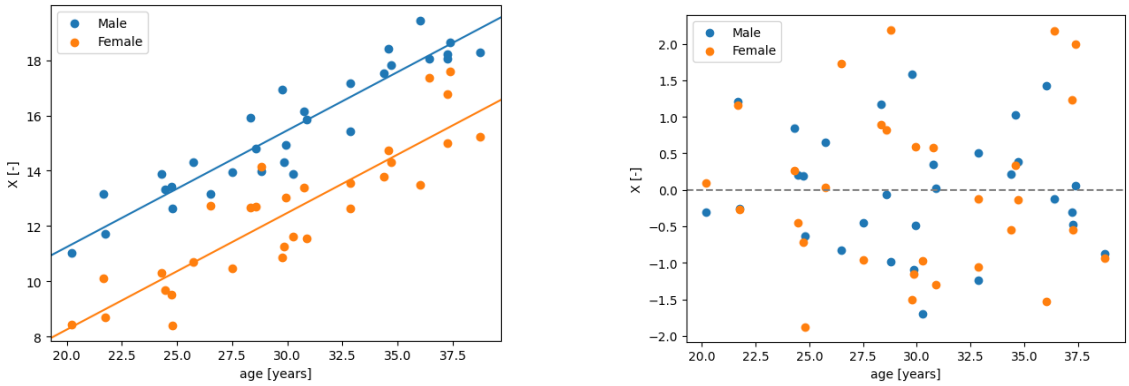

Figure 1.3: Visual sketch of the orthogonalization procedure using a linear model. Left: fitted model of the dependent variable as function of sex and age, to obtain the residuals. Right: distribution of the residuals (age and sex-corrected variable X) as function of sex and age.

1.6.8 Evaluation

Predictor performance is evaluated and compared using scoring metrics. For classifiers, these metrics are derived from the confusion matrix (Figure 1.4) and chosen with respect to specific goals or use cases. The matrix summarizes the interaction of prediction and ground truth: true positives (TP) and true negatives (TN) for correct classification, false positives (FP) and false negatives (FN) for incorrect classification. The most common and straightforward metric is accuracy (or its complement, error).

Accuracy (classification) is the proportion of correctly classified examples to all examples.

$$ACC = \frac{TP + TN}{P + N} \tag{1.20}$$

Coefficient of determination R^2 (regression) expresses how much variance is captured by a regression model in comparison to the null model (simple mean of data). It is computed as

$$R^2 = 1 - \frac{SS_{\text{residual}}}{SS_{\text{total}}} \tag{1.21}$$

where SS_{residual} is the residual sum of squares (for evaluated model) and SS_{total} is the total sum of squares (for the null model), which is equivalent to the variance of the data. As R^2 is observed to increase just with adding predictors into a model, and adjusted variant has been devised to compensate for this behavior:

$$adjR^2 = 1 - (1 - R^2)\frac{n - 1}{n - p - 1} \tag{1.22}$$

Where n is the sample size and p is the number of predictors in the evaluated model.

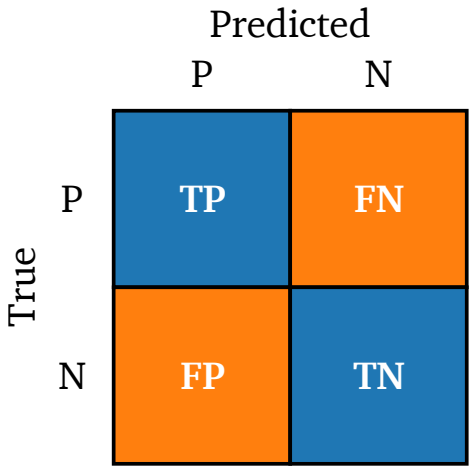

Figure 1.4: The confusion matrix.

1.6.9 Validation

Validating a model means testing how the model fares on predicting unseen data that weren't in the training set. The key aspect of model validation is having separate data for training and testing the model. The trivial case to satisfy this condition would be *holdout* validation: splitting the dataset into two non-overlapping subsets, training on one of them, and testing predictions on the other. Conventionally, the training set is larger than the testing one, often in the 70:30 ratio. This is a valid approach; however, it does not provide a measure of uncertainty and is quite sensitive to how the examples in both sets are distributed. The solution to this are resampling methods such as *cross-validation*.

1.6.10 Cross-validation

Cross-validation (CV) alleviates the aforementioned issue by dividing the training set into smaller subsets. One subset is put away, and the training runs on the remaining subsets. The classification score is then computed on the held-out validation subset. This yields a set of training scores, which are then averaged. Other statistics, such as evaluation time, are also often computed, providing additional information about the classifier. CV is conventionally implemented as k-fold CV, dividing the training set into k sets of roughly the same size, using $k - 1$ sets as the training set and the remaining set as the validation set. The extreme case of k-fold CV is called *leave-one-out*, where the validation set consists of only one example. *Stratified* k-fold CV is often applied when there is a noticeable class imbalance in the sample. It sorts training examples into strata corresponding to their class label, attempting to preserve in each fold the class ratio of the dataset. *Repeated* CV takes any of the former defined procedures and repeats them a number of times, shuffling the dataset each time, again averaging over folds and repetition. It is intended to be more robust to dataset permutation. A variant popularized when Dietterich (1998) introduced

it for model comparison is *5x2cv*, which shuffles the dataset five times, making the metric less sensitive to dataset ordering, while still fitting the model 10 times like in often used 10-fold CV. *Cross-validated prediction* is a sort of model diagnostic prediction, where instead of scores we are interested in the predicted values in folds, which are then concatenated into one single vector for the whole predicted set. This vector is comprised of predictions from k different models and, as such, is used for comparing different algorithms rather than the model they produced. To contrast this, when we use the holdout method, train the two models once on the same training set, and compare predictions from the same testing set, we are performing a model comparison.

1.6.11 Data leakage

Importantly, to avoid data leakage, the validation examples ought not to be included in fitting transformations such as PCA or confounder orthogonalization. Data leakage refers to properties of testing examples being projected into the estimator before being classified. For example, principal components could be heavily influenced by the validation examples in smaller samples (this might be less of an issue in LOOCV, but the principle holds). So, the transformation is fitted on the training set (which is then also transformed), and the validation set is only transformed by the fitted model.

1.6.12 Grid search

Grid search is a technique for optimizing estimator hyperparameters. It cross-validates the classifier for each combination of considered hyperparameters, and the parameter combination with the best mean score is selected for the job. This way, parameters like regularization strength or optimal number of (PCA) components can be estimated on the training set to achieve a good score while preventing overfitting to the training set. Moreover, multiple classifiers with different principles can be hyperparameter-optimized on the training set and compared in their optimal tunings.

1.6.13 Comparing classifiers

McNemar's statistical test is used in situations where we want to compare two paired samples of nominal data. As such, it is often used for pairwise comparison of diagnostic tests, which includes classification models. The usage of McNemar's test for comparing the performance of supervised models is detailed by Dietterich (ibid.). First, a contingency table counting concordant and discordant pairs is constructed (Table 1.2). In the case of classifiers, pairs are counted that were concordant and discordant in terms of correct classification of the training example. Following that, the test statistic is

Table 1.2: Contingency table of McNemar's test between classifiers A and B.

	A is correct	A is incorrect
B is correct	a	b
B is incorrect	c	d

$$\chi^2 = \frac{(|b-c|-1)^2}{b+c} \tag{1.23}$$

The test statistic follows the χ^2 distribution with one degree of freedom. The 1 in the numerator of eq. (1.23) is added for "continuity correction" because of the fact that the statistic is discrete, but the distribution is continuous. We can see that the test takes into account only the discordance between the correctness of classification of the algorithms. The hypotheses for this test are the following:

$$H_0 : p_b = p_c \tag{1.24}$$

$$H_1 : p_b \neq p_c \tag{1.25}$$

The null hypothesis expects the counts b and c to be equal to their average $(b+c)/2$. It is not dependent on the concordance counts a and d.

Cochran's Q-test is an omnibus test that assesses differences between more than two models of classification. It can be seen as an extension of McNemar's test, akin to analysis of variance being the omnibus alternative to pairwise t-tests between groups. Given the ground truth class labels and corresponding predictions for each classifier, we can construct a table that records the correct predictions of subjects in rows per each classifier (example in Table 1.3). Following from this, the Q statistic is computed by the formula

Table 1.3: Table of Cochran's Q test between classifiers A, B, and C.

	Classifier A	Classifier B	Classifier C
Subject 1	I_{11}	I_{12}	I_{13}
Subject 2	I_{21}	I_{22}	I_{23}
Subject 3	I_{31}	I_{32}	I_{33}
⋮	⋮	⋮	⋮
Subject n	I_{n1}	I_{n2}	I_{n3}

$$Q = k(k-1)\frac{\sum_{j=1}^{k} I_{\cdot j} - N/k}{\sum_{i=1}^{n} I_{i\cdot}(k - I_{i\cdot})}, \tag{1.26}$$

where I_{ij} denotes an binary variable indicating the correctness of the classification, k is the number of classifiers compared, n is the number of subjects classified, $I_{\cdot j}$ and $I_{\boldsymbol{i}\cdot}$

denote the sum of columns and rows respectively, and N is the sum of all ones (correct classifications) in the table. The hypotheses for this test are following:

$$H_0 : \text{The treatments are equally effective.} \rightarrow ACC_1 = ACC_2 = \ldots \quad (1.27)$$

$$H_1 : \text{There is a difference in effectiveness between some treatments.} \quad (1.28)$$

Dietterich (1998) reviewed several ways of comparing algorithms, and presented a new approach. The *5x2cv paired t-test* performs repeated 2-fold cross-validation. The improvement lies in the 2-fold part, as the train and test sets are independent (the repetitions still are not, as the sets are shared across repetitions). This test has been shown to have similar Type I error to McNemar's test, while having greater power (ibid., pp. 20–23). While the former two tests are specific to classification, 5x2cv can be used for regression as well.

1.6.14 Matters of size

The number of training examples influences the behavior of machine learning predictors. Here, we assess how this role sample size depends on the modalities studied.

Learning curves are used to analyze the behavior of model performance with varying training sample sizes. Learning curves are conventionally constructed by repeatedly subsampling a dataset at different sample sizes and cross-validating the inspected model on these subsamples. Cortes et al. (1993) define learning curves as "the expectation value of the test and training errors as a function of the training set size l."

In our approach, for each sample size studied, we subsampled our data to this size (without replacement; assuming balanced groups) and cross-validated a classifier on this subsample (10-fold stratified CV). This was repeated for a number of times (here the number of repeats was 100) for each sample size and averaged.

1.7 Fusions

When referring to multimodal fusion, we talk about a specific form of feature engineering which from several feature sets generates a single fused one. This has prompted nomenclature naming the strategies in the engineering pipeline as described by S.-C. Huang et al. (2020). Note that in this work, we use *fusion* and *integration* interchangeably.

Early fusion joins features from all modalities into a single vector to be input into a machine learning model.

Late fusion feeds each modality into a separate machine learning model, joining the results from these models into a feature vector do be used in learning.

Joint fusion aims to learn a joint representation space for the input modalities by optimizing the submodels, which output the intermediate features, along with the final model.

1.8 Survey of schizophrenia prediction

Individual prediction of schizophrenia has been a rising topic in research for the last twenty years. Just in the past decade, the technology and methodology of the MRI have developed substantially and, as such, provided greater resolution, reliability, and accessibility. Also owing to these developments, sample size, a problematic aspect of studies, has grown from tens of subjects in total (e.g. Arbabshirani et al. (2013)) to hundreds of subjects in each group in multi-site experiments (e.g. Vieira et al. (2020), H. Cao et al. (2022), and Dwyer et al. (2023)).

1.8.1 Structural imaging

As the oldest modality, brain sMRI of patients with schizophrenia have been studied in group analyses for decades. The structure of the brain is more subtly affected by schizophrenia, as suggested by more than 40 years of research, wherein group analyses on the modality have not yet found conclusive clinical markers of the illness (Fusar-Poli and Meyer-Lindenberg 2016). Additionally, this research alone does not provide evidence of applicability in individual prediction. Nevertheless, the results obtained do serve as a foundation for conducting prediction research and are useful for comparison with results achieved therein.

Chin et al. (2018) used a voxel-wise support vector machines (SVM) on on VBM images. A form of spatial and anatomical regularization has been used. Spatial regularization intends to attenuate the effects of local variations, and anatomical regularization suppresses inter-regional interactions, and these regularizations were implemented as an SVM kernel. Additionally, a sequential region of interest (ROI) selection was included. There, voxels from an initial ROI are included. Similarly to forward step-wise feature selection, for each remaining ROI in turn, the ROI that improves the accuracy the most is selected and its voxels included. This procedure was performed for each ROI (from the Harvard-Oxford cortical and subcortical atlases) as a starting point, yielding 64 feature selection trajectories. Counterbalancing the accuracy and computational load of the procedure, 7 ROIs were selected for the final classifier. This approach yielded a strong test accuracy of 89.4%. However, some limitations to this study include an unbalanced dataset

(roughly 2:1 SZ to HC) and, on average, long-term patients with medication, so the question of applicability on first-episode psychosis drug-naive patients stands.

Yamamoto et al. (2020) also performed classification on whole-brain voxel-based morphometry. Collecting subjects from two sites, an SVM pipeline has been trained on each site and tested for generalization on the other, with both classifiers attaining a cross-site consistent accuracy of slightly above 70%. While this approach sacrifices training size for the individual models, it circumvents the need to harmonize the entire dataset for a single model and shows the generalizability of a single model architecture trained on different sites. The authors also listed regions of interest most contributing to the accuracy of the model. To compare with Chin et al. (2018), while there were some common regions, such as occipital fusiform gyrus or middle frontal gyrus, most regions differed either completely, or the region was a different part of the same structure, e.g. posterior as opposed to the anterior part of the superior temporal gyrus.

Anatomical subgroups have been identified among the patient population by Chand et al. (2020), joining the ranks of studies attempting to stratify the population affected by the disorder. The subgroups are defined as SG1 - lower brain volume and SG2 - higher striatal volume, and they were identified in data of predominantly longtime patients. In a follow-up study, these subgroups were subject to further investigation by Dwyer et al. (2023). In this study, the anatomical subgroups were investigated in both healthy control (HC) and schizophrenia patient (SZ) by classifying patients into one of four categories: SG1, SG2, SG1 + SG2, None. Proportions of controls and patients have been analyzed in each category. first-episode psychosis (FEP) participants were significantly more represented in the SG1 and SG2 subgroups, supporting the hypothesis of early changes in structural images of first-episode patients. Furthermore, in the SG2, a greater number of patients with symptom remission were present. Apart from anatomical stratification, these subgroups then provide a hint at future functional outcome and deserve further investigation in treatment trials. While there was a considerable fraction FEP patients not assigned to either SG1 or SG2, meaning that these two subgroups may present a small sample of such anatomical variations, this method represents a promising approach in subtyping schizophrenia.

In addition to the stratification of patients within the schizophrenia diagnosis, comparing schizophrenia with other mood and psychotic disorders is an important branch of research since schizophrenia is such a heterogeneous disorder. Salvador, Radua, et al. (2017) included a cohort of patients with bipolar patient (BP) group in addition to the HC and SZ groups. Among the considered sMRI features, the largest classification accuracy (75%) was achieved with VBM features between the HC and SZ groups (regardless of predictor used). The accuracy between BP and HC (63%) or BP and SZ (62%) were weaker, as feature values BP participants were scattered between those of the other two groups.

1.8.2 Functional imaging

So far, the literature on unimodal prediction points to functional MRI as consistently the most powerful modality, with classification accuracy on resting-state fMRI features reaching upwards of 80% (Steardo et al. 2020). While this line of research identifies the predictive potential of fMRI, even the best-performing classifiers will have to undergo proper validation and scrutiny before any clinical use. For example, in Arbabshirani et al. (2013), the best-performing classifier provides an accuracy of 96% on functional network connectivity features. However, this number is severely inflated by the fact that the training sample is only 56 subjects (28 HC + 28 SZ). A sample size this small does not capture the heterogeneity of the condition, which usually leads to over-optimistic results.

In addition to common confounders such as age, sex, or recording site, an important role is played by medication. Antipsychotic medication has both desirable influence and side effects projecting into the data measured. As such, the application of models trained on medicated patients can have limited application on incoming patients with first-episode psychosis. Kalmady et al. (2019) is one of the first studies recruiting a cohort of fully drug-naive patients for model training. Regional and connectivity measures from fMRI have been computed for multiple brain parcellations, and a probabilistic classifier has been trained on these features. The ensemble classifier then consisted of averaging the probabilities of all the classifiers (soft voting), where the class is assigned based on the averaged probability. This model achieved an accuracy of 87%. While this study used around 3$\times$ more subjects than Arbabshirani et al. (2013) and includes unmedicated patients, there is no account on the duration of untreated psychosis, and females are slightly unrepresented (around 1/3 in each group). The soft voting principle is not unique to this study and is suitable for integrating the input of classifiers trained on multiple modalities.

Following a group-level study (Hengyi Cao et al. 2018) identifying patterns of increased connectivity in the cerebello-thalamo-cortical network as a neural marker for (independent of clinical/behavioral presentation) psychosis, H. Cao et al. (2022) focus on connectivity features in this network as a means of individual prediction. Acquiring data from a relatively large cohort of participants (179 HC; 214 SZ, all of which were yet untreated FEP patients). A fraction of the patients (62 participants) underwent at least one follow-up visit in the 12 or 24 months after the initial visit, providing longitudinal data for the prediction of outcome. The results found that hyper-connectivity in the CTC network predicted lesser remission of negative symptoms as indicated by the negative subscale of PANSS. Negative symptoms are generally more functionally impairing and less likely to be improved by treatment than positive symptoms. Finding a biomarker indicating the persistence of negative symptoms would help prioritize their management to potentially improve the illness course.

1.8.3 Diffusion tensor imaging

Diffusion tensor imaging is the youngest of the three modalities discussed. It is also more sensitive to noise and artifacts, requiring stricter quality assurance and (largely manual) quality control (Liu et al. 2010). Moreover, owing to more degrees of freedom in the preprocessing and computational steps (Oldham et al. 2020), there is more diversity in the pipelines, making it less viable to compare studies or perform meta-analyses on large samples if the raw data is not provided. For these reasons, although there exists a line of evidence suggesting group-level changes in white matter integrity present in early-stage schizophrenia (Samartzis et al. 2014), research on the prediction on DTI features in schizophrenia is a small niche. Nevertheless, there have been efforts to use DTI in prediction.

Mikolas et al. (2018) surveyed a cohort of first-episode schizophrenia (FES) patients in a medium-sized study (77 participants in each group, age/sex-matched). A linear SVM on voxel-wise fractional anisotropy (FA) achieved a significant but weaker accuracy of around 62%. While this result is not strong in itself for schizophrenia prediction, the study also explores the influence of medication doses and symptom scores on classification accuracy, showing that such variables may not be *the* factors negatively impacting the predictive power of DTI measures. On the other hand, classification probability estimates (obtained from SVM via Platt scaling (Platt 1999)) were not significantly correlated with any clinical variable. Furthermore, group analyses of fractional anisotropy showed clusters of significant difference between the groups. While these clusters did overlap with regions contributing to prediction accuracy, these results demonstrate the discrepancy between group analyses and individual prediction.

In an attempt to help the prediction performance of the modality, Elad et al. (2021) used normative modeling to compute age and sex-adjusted z-scores of diffusion measures on a large harmonized multisite dataset containing data from schizophrenia-spectrum patients (512 HC, 601 SZ). These z-scores, representing deviations from the normative model, were then used in classification. Three diffusion MRI measures were computed: fractional anisotropy (FA), fractional anisotropy in the tissue compartment (FAt), and fractional volume of the free-water compartment (FW). These were then averaged over white matter skeleton regions. The raw values of fractional anisotropy in individual regions provided weak predictive performance in terms of AUC, and the z-scored FA provided only mild improvement. The best performance for this measure was achieved by combining all regions in prediction (AUC = .67). FAt and FW features generally outperformed FA when individual regions were used for prediction; however, the best overall classifier was constructed by combining both FAt and FW features over all regions (AUC = 0.726). Overall, the normative approach comprehensively improved the predictive power of the studied feature. In the baseline, normative modeling serves as a means of controlling for con-

founders and thus is suitable for large datasets as used here, though this study still might have suffered from diverse durations of illness in patients.

Another approach is shown by Tønnesen et al. (2020), where a hypothetical brain age is modeled on DTI data. A regressive model is trained strictly on a healthy population, with brain data as a predictor and age as a target variable. In this way, a model of a normative brain age is constructed. This model is then applied to the patient population, with the difference between the true and predicted brain age being the deviation from normal aging. The study has shown that in this model, the brain age of patients is largely overestimated, suggesting some influence of schizophrenia on normal brain development. While the concept of brain-age prediction is an interesting addition to the schizophrenia prediction toolbox, there are several limitations to the approach. This method is essentially the opposite way of conventional normative modeling, with the disadvantage that the brain variables are much more complex and varied than common demographic factors. While age and sex are common variables with known and stable distributions, the brain-age model undoubtedly has more degrees of freedom, such as choice of modality, preprocessing, brain parcellation, feature combinations, and much more. This makes the predicted brain age more dependent on the construction and assumptions of the model rather than actual brain development. Also, the actual definition of brain age itself might not be straightforward. Taken together, while this is an interesting approach and might bear fruit in some cases, it is recommended to proceed with caution, as the researcher's degrees of freedom are vast.

1.8.4 Multimodal imaging

Unimodal analyses have been conducted for quite some time now, and throughout, a leap was made in improving methodology and understanding each modality's limitations and strengths. Recent years have shown a large increase in the number of multimodal studies investigating to what extent multimodal fusion can overcome the limitations of individual modalities. Multimodal fusion or integration is the usage of at least two sources of neuroimaging contrast, such as different MRI contrasts, functional MRI together with EEG recorded simultaneously, and so on. Generally, multimodal studies provide unimodal analyses of the inspected dataset for comparison with the multimodal results.

Salvador, Canales-Rodríguez, et al. (2019) demonstrate a simple multimodal approach. Measuring resting-state, task-based fMRI, and structural MRI features, select classifiers were first applied to every modality individually. The largest mean accuracy of 84% was reached with lasso regularized logistic regression on activation maps of the 2back task in fMRI. Then, a redundancy assessment was conducted between all feature sets to find out if feature set 1 provides additional predictive information with respect to feature set 2 and vice versa. The multimodal integration was performed in multiple ways. First,

classification probabilities for each modality were collected. Functions such as the mean (soft-voting) or max were applied to this collection, with the output probability from this function determining the class label. The probabilities were also transformed into log-likelihood and input as features to a classifier in a two-step classification scheme. Another way was to use the classifiers themselves as feature selection procedures - for instance, using only voxels with nonzero coefficients from the lasso classifier, then again training a final classifier only on these features. Finally, a neural network was devised, combining the brain maps in convolutional and fully connected layers. This enumeration is to show that the pipelines in multimodal analyses can be diverse and also quite arbitrary. In the end, the best multimodal classifier (two-step ridge classifier) provided a slightly better accuracy (87%) than the best unimodal classifier (84%). While hopes are that combining more types of data will lead to improved performance, it is not to be applied naively, as blindly combining too many features may even worsen the accuracy of a classifier.

Ambrosen et al. (2020) combine cognitive, EEG, and structural MRI data to build a baseline classifier and also a predictor of short/long-term functional outcome via PANSS. The patient data came from three consecutive cohorts measured over the span of 16 years (1998-2014) as a part of medication treatment studies; however, the participants were measured before the medication protocol started, providing a large dataset from antipsychotic-naive FES patients (138 participants). One hundred fifty-one controls were recruited for this study to match demographic values in the patient group. The machine learning approach was divided *single* and *ensemble* models to predict the short-term (binary HC/SZ) and long-term (PANSS regression target). The single approach consisted of testing common linear and nonlinear models individually; the ensemble approach was implemented by an AutoML framework (auto-sklearn), which searches a vast model and parameter space to compose a meta-classifier with improved performance. Furthermore, to avoid bias in selecting models on real data, a synthetic dataset on which algorithm choices were made was simulated from properties of the real dataset, and the resulting models with parameter constraints were trained and validated on the real dataset. While this overall pipeline is quite complex, the balanced accuracy of the diagnostic classification task reached 63.8% for the single-model approach and 64.2% for the ensemble approach, with the classification more influenced by cognitive tests rather than neuroimaging data. For the long-term classification task, both approaches approached 50% balanced accuracy (chance). These results suggest that complicated approaches do not always improve performance.

Rahaman et al. (2021) integrated structural and functional MRI and genomic data in a deep-learning approach. Functional network connectivity is extracted from fMRI, group independent component analysis loadings from sMRI, and single nucleotide polymorphisms (SNP) from genomics. Each feature set then goes through a specialized neural subnetwork, which outputs a latent representation or embedding. These embeddings are

then fused together via weighted concatenation into a single vector and finally forwarded into a deep neural model that predicts the diagnosis. In addition to the full multimodal model, unimodal and pairwise multimodal models have been trained for comparison. The full model has achieved the greatest accuracy (88%), compared to the best unimodal model based on fMRI (81%) and a partial multimodal model based on sMRI and fMRI (83%). The study also investigates the explainability of the deep model by working with saliency maps, which indicate the importance of features in the correct prediction of a sample. Identified features were further statistically analyzed and showed some significant differences between groups. Furthermore, genes were identified from the salient SNP data, and genetic pathway analysis has been conducted on them. Such an approach could potentially lead to identifying new genetic targets in drug development or repurposing.

1.8.5 Remarks

So far, multimodal models have achieved moderate margins in classification performance over unimodal models in the litarature. Moreover, multiple types of data do not only improve performance but can also provide new insights. While caution should be exercised in framework design – degrees of freedom are numerous in multimodality, and too much complexity may even hinder performance – the multimodal approach provides diverse avenues in research going forward. It is, however, key to keep track of the limitations that multimdality brings, so the burdens wont overweigh the advantages.

Higher-dimensional data that are innate to multimodal analyses require more training examples to be better comprehended by both researchers and the learning algorithms. With larger samples, more heterogeneity among patients can be captured, so machine learning models can make more powerful inferences about the population, and so multi-site studies aggregating data from more recording centers are becoming the norm. This comes with its own class of challenges, as multiple sites introduce variability in recordings partly due to differences in technical equipment and demographics on site. The larger and more diverse the sample, the more need is there to account for the effects of confounder variables. Confounder correction is commonly done by age/sex matching of the groups or regression decomposition of the predictor variables by the confounders. Newer approaches include normative modeling, where brains are adjusted by the age/sex group (Dinga et al. 2021) or specialized methods like ComBat multi-site harmonization (Johnson, Li, and Rabinovic 2007).

Chapter 2

Prepare Dataset

2.1 Data provenance

The dataset used in this thesis comes from the data collected as part of the ESO project. The ESO project is a large-scale effort orchestrated by National Institute of Mental Health (NUDZ) aiming to predict the treatment trajectory and outcome of schizophrenia patients. Patients are recruited from the Bohemia surveillance area shortly after their first episode. Undergoing three visits to the study site in total, a longitudinal multimodal dataset is formed. In addition to all modalities recorded on MRI scanners, patients go through a battery of clinical, laboratory, and neurocognitive examinations (Hlinka 2020). The healthy control sample underwent a similar protocol sans patient-specific procedures (such as the PANSS interview, aiming to establish the severity of schizophrenia in diagnosed patients).

NUDZ collects and governs data in the Hyper Database of Recognizable Associations (HYDRA). In this environment, raw data is stored, further processed by pipelines and derived into features. Authorized persons are allowed to filter and export data for analyses. We have requested data from the ESO project for our analyses and acquired several files that are used to assemble the dataset.

2.2 Scanners

The data measured come from two sites: Institute of Clinical and Experimental Medicine and National Institute of Mental Health. The IKEM site has served as a scanning site with a 3 Tesla scanner (Siemens Magnetom Trio 3 T) until another 3 Tesla scanner (Siemens Magnetom Prisma 3 T) was installed at NUDZ, and the measurements relocated there. Patients who made their first visit to the IKEM site follow up on the same site; any newly recruited patients are measured exclusively at NUDZ.

IKEM parameters

sMRI T1 structural image acquired using the magnetization prepared rapid acquisition gradient echo (MPRAGE) sequence with inversion time (TI): 900 ms, repetition time (TR): 2300 ms, echo time (TE): 4.63 ms, flip angle: 10°, 1 average, matrix: 256 × 256 × 224, voxel size: 1 × 1 × 1 mm^3, bandwidth: 130 Hz/pixel, GRAPPA acceleration factor: 2 in phase-encoding direction, reference lines: 32, prescan normalize: on, elliptical filter: on, raw filter: off, acquisition time: 5:30 minutes.

dwMRI Acquired using a Spin-Echo EPI sequence with TR: 8300 ms, TE: 84 ms, matrix: 112 × 128, voxel size: 2 × 2 × 2 mm^3, b-value: 0 and 900 s/mm^2 in 30 diffusion gradient directions, 2 averages, bandwidth: 1502 Hz/pixel, GRAPPA acceleration factor: 2 in phase-encoding direction, reference lines: 24, prescan normalize: off, elliptical filter: off, raw filter: on (intensity: weak), acquisition time: 9:01 minutes.

fMRI Acquired using T2-weighted echo-planar imaging (EPI) with blood oxygenation level-dependent (BOLD) contrast using SENSE imaging. GE-EPIs with TR: 2000 ms, TE: 30 ms, flip angle: 70°, 35 axial slices acquired continuously in sequential decreasing order covering the entire cerebrum, voxel size: 3 × 3 × 3 mm, slice dimensions: 48 × 64 voxels. 400 functional volumes were used for the analysis. A three-dimensional high-resolution MPRAGE T1-weighted image with TR: 2300 ms, TE: 4.63 ms, flip angle: 10°, voxel size: 1 × 1 × 1 mm covering the entire brain was acquired at the beginning of the scanning session and used for anatomical reference.

NUDZ parameters

sMRI T1 structural image acquired at the National Centre of Mental Health in Klecany, Czech Republic, at the National Institute of Mental Health using Siemens MAGNETOM Prisma 3T. Acquisition parameters of T1-weighted images using MPRAGE sequence: 240 scans, voxel size: 0.7 × 0.7 × 0.7 mm, repetition time (TR): 2400 ms, echo time (TE): 2.34 ms, inversion time (TI): 1000 ms, flip angle: 8°, field of view (FOV): 224 mm × 224 mm, acquisition matrix: 320 × 320.

dwMRI Acquired using a Spin-Echo EPI sequence with TR: 3000 ms, TE: 82 ms, voxel size: 2 × 2 × 2 mm^3, matrix: 100 × 100, FOV: 200 mm × 200 mm, b-value: 0, 1000, and 3000 s/mm^2 in 74 diffusion gradient directions, 6 averages, bandwidth: 1515 Hz/pixel, GRAPPA acceleration factor: 3 in phase-encoding direction, acquisition time: 5:15 minutes.

fMRI, version 1 Acquired using the T2-weighted (T2w) gradient echo-planar imaging (GR-EPI) sequence sensitive to the blood oxygenation level-dependent (BOLD) signal with TR: 2000 ms, TE: 30 ms, flip angle: 70°, voxel size: 3 × 3 × 3 mm^3, field of

view (FOV): 192 mm × 192 mm, matrix size: 64 × 64, each volume with 37 axial slices (slice order: alternating increasing), 300 volumes in total.

fMRI, version 2 Acquired using the T2-weighted (T2w) gradient echo-planar imaging (GR-EPI) sequence sensitive to the blood oxygenation level-dependent (BOLD) signal with TR: 1000 ms, TE: 30 ms, flip angle: 52°, voxel size: 3 × 3 × 3 mm^3, field of view (FOV): 222 mm × 222 mm, matrix size: 74 × 74, each volume with 60 axial slices (multiband sequence with factor 4), 400 volumes in total.

2.3 Features

Table 2.1 lists all MRI features we use.

Table 2.1: MRI modalities and features

MRI Modality	Techniques
Structural MRI	Voxel-based morphometry
Functional MRI	Functional connectivity Amplitude of low-frequency fluctuations
Diffusion-weighted MRI	Structural connectivity Fractional anisotropy Axial diffusivity Mean diffusivity Radial diffusivity

2.3.1 Structural MRI

VBM has been already precomputed in HYDRA in the 90 gray matter AAL regions.

Voxel-based morphometry is an approach for detecting statistical differences in voxels of aligned structural images between groups of subjects. Here, instead of statistical group comparison, we use the interim result of voxel-wise gray matter density estimate for each subject, further averaged across regions to provide suitable number of robust features.

2.3.2 Functional MRI

For this modality, we worked with preprocessed timeseries of 400 timepoints in 90 AAL gray matter regions for all participants. All time series were then orthogonalized against signal of cerebrospinal fluid, white matter, and 12 motion parameters (six original parameters and their temporal derivatives), linearly detrended and bandpass filtered using

the Butterworth filter with a window of 0.008 - 0.09 Hz, so the signal only contains low frequency fluctuations (Hlinka et al. 2024).

Functional connectivity is defined as statistical dependence (here quantified by Pearson's correlation) of timeseries from two spatially separated regions. Thus, we computed the coefficient for each pair for the 90 AAL regions into a connectivity matrix. This matrix is symmetric, and so to transform it into a feature vector, we only need to extract the lower or upper triangle, sans diagonal (Figure 2.1).

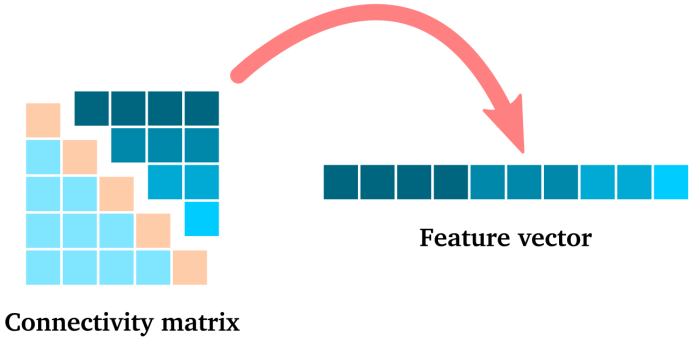

Figure 2.1: Extracting symmetric matrix into a feature vector.

Amplitude of low frequency fluctuations is defined as root mean square of the fMRI BOLD timeseries filtered to only contain low frequency fluctuations (Yang et al. 2007).

2.3.3 Diffusion-weighted MRI

The dwMRI features have been already precomputed in HYDRA. Structural connectivity has been computed between the 90 AAL regions, diffusion tensor metrics have been averaged across the 50 regions of the JHU atlas. Because of missing values in all of the tensor metrics except fractional anistropy (FA), we've excluded features with the missing values and left in 28 features each for the axial diffusivity (AD), mean diffusivity (MD), and radial diffusivity (RD) metrics.

Structural connectivity is computed via methods of probablistic tractography, which output white matter tracks. In the pipeline that generated our structural connectivity matrices, the connectivity between ROI1 and ROI2 is estimated as follows: fibers seeded (starting) in any voxel of ROI1 and passing through any voxel of ROI2 are counted. This number is then normalized by the number of voxels in the seed ROI and also by the number of fibers seeded per voxel (5000). The resulting number is then interpretable as a *connectivity probability* (Škoch et al. 2022). The connectivity matrices are not symmetric in general, but they are "almost" symmetric, so to enforce

this pragmatic requirement, the matrices are symmetrized

$$S_{sym} = (S_{assym} + S_{assym}^T)/2 \tag{2.1}$$

Axial diffusivity measures diffusivity along the strongest direction:

$$AD = \lambda_1 \tag{2.2}$$

Mean diffusivity measures mean diffusivity along all three axes of the ellipsoid:

$$MD = \hat{\lambda} = \frac{\lambda_1 + \lambda_2 + \lambda_3}{3} \tag{2.3}$$

Radial diffusivity measures diffusivity in the diretions perpendicular to the main diffusion direction (along an axon):

$$RD = \frac{\lambda_2 + \lambda_3}{2} \tag{2.4}$$

Fractional anisotropy expresses the degree of anisotropy in the voxel. If the ellipsoid is a sphere, it means the water molecules can move in any direction and the value is zero. If the ellipsoid collapses to a line, the motion of the molecules is restricted only to this direction, and the value is one.

$$FA = \sqrt{\frac{3}{2}} \frac{\sqrt{(\lambda_1 - \hat{\lambda})^2 + (\lambda_2 - \hat{\lambda})^2 + (\lambda_3 - \hat{\lambda})^2}}{\sqrt{\lambda_1^2 + \lambda_2^2 + \lambda_3^2}} \tag{2.5}$$

$$\tag{2.6}$$

2.4 Covariates and clinicals

Along with the MRI modalities, we have processed the covariates that were used for controlling for confounders, as well as clinical PANSS scores and diagnosis labels.

Patient Indication of whether the participant is in the HC group (0) or the SZ group (1).

Sex Biological sex of the participant.

Visit age Age of participant on the date of the first visit.

PANSS positive measures positive symptom severity.

PANSS negative measures negative symptom severity.

PANSS global measures the severity of global psychopathology.

PANSS total is a sum of the three subscales.

2.5 Dataset disparity

Due to data governance and prioritization in HYDRA and the running length of the ESO project itself, there are differences in final training (IKEM) and testing (NUDZ) datasets. Apart from some newer processing pipelines in NUDZ dataset, the structural connectivity has not been computed as of the writing of this thesis. From the diffusion metrics, only the FA features will be present since the other feature sets have high proportions of missing data. As such, these features will not be included in validation (Chapter 4), and the comparison of the two datasets is presented in Table 2.2. Summary statistics for both datasets are shown in Table 2.3.

Table 2.2: Training and testing dataset feature set availability.

	FC	ALFF	VBM	SC	FA	AD	MD	RD
IKEM (training)	✓	✓	✓	✓	✓	✓	✓	✗
NUDZ (validation)	✓	✓	✓	✗	✓	✗	✗	✗

Table 2.3: Training and testing dataset summary statistics.

	Sample size	M:F	HC:SZ	Visit age
IKEM (training)	161	83:78	78:83	27.8 $\pm$ 6.5 years
NUDZ (validation)	204	111:93	76:128	28.3 $\pm$ 7.7 years

The effects of different preprocessing pipelines can be demonstrated in difference of accuracy for these pipelines, see an example of classification accuracy from functional connectivity features with several different preprocessing in Figure 2.2. For more detailed discussion, see the paper by Hlinka et al. (2024).

2.6 Feature vector

All of the features per participant were composed into a single feature vector, which can be sliced to contain only the feature sets needed for a concrete analysis. The subvectors for each feature sets will have lengths as shown in Table 2.4.

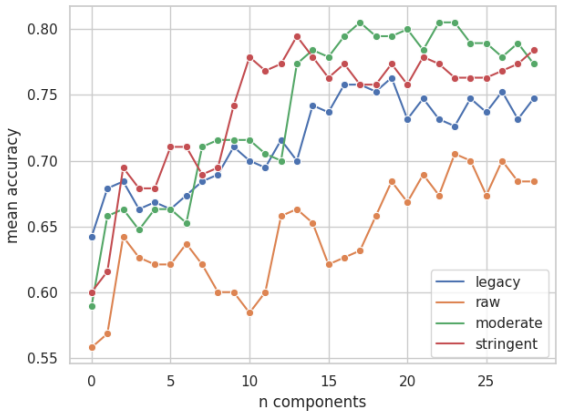

Figure 2.2: Dependence of linear classifier accuracy on the preprocessing pipeline and number of components used.

Table 2.4: Training and validation dataset feature vector lengths.

	FC	ALFF	VBM	SC	FA	AD	MD	RD	total
IKEM (training)	4005	90	90	4005	50	28	28	28	8324
NUDZ (validation)	4005	90	90	–	50	–	–	–	4235

2.7 Implementation notes

We have implemented the entirety of our analyses in the Python language. Our modules for data loading and the bulk of our work were done within the scikit-learn framework, which provides an object-oriented interface to data processing, classical machine learning models, and ways to compose them. We have developed our own models for transforming the data in this framework, namely the `Orthogonalizer` class, which, given a table of covariates for each training example, regresses out these variables from the training data; the `SelectorTransformer` is used for filtering columns of the dataset to a subset of modalities. In this way, we can supply the entire dataset to the estimator without having to filter it manually each time, as the modality parameter is easily set. This class is dependent on the way we name the columns in our dataset – if we want to limit it to e.g. functional connectivity (FC), all such columns in the dataset must contain the substring "FC". As such, the column names in our dataset follow this convention:

$$\{\text{modality acronym}\}\_\{\text{atlas acronym}\}\_\{\text{feature name}\} \tag{2.7}$$

For example, the functional connectivity between AAL regions of left precentral gyrus and right angular gyrus will be labeled `FC_AAL_Precentral_L_vs_Angular_L`. This way, even if there would be an ambiguity in the name of the feature itself, we are sure to select functional connectivity features by matching the pattern "FC_AAL_" with the function `select_modality(data, "FC_AAL_")`.

The entire dataset assembly pipeline is located in the dedicated scripts

- `prepare_dataset_ikem.py`
- `prepare_dataset_nudz.py`

which take IKEM/NUDZ site-specific files, perform cleaning, feature derivation, and data wrangling on each, align the tables for both modalities and covariates by the participant (HYDRA) ID, and save these tables into separate `parquet` files. Parquet is a columnar table data format that allows for efficient storage and preserves column data types. The `load_dataset(ds_name)` is a utility function that returns the contents of these files as `pandas` dataframes. When we want to perform joint analyses, such as validation, between the datasets, we have to align them on the column axis to include only features present in NUDZ, as it is a feature subset of IKEM.

Chapter 3

Explore & Design

In this chapter, we perform explorations on the IKEM dataset. We will get an idea about the data and evaluate the possibilities of prediction on multimodal data. In this process, we will design a number of pipelines, some of which will be tested later on the NUDZ dataset.

3.1 Demographics

Of the healthy group, 42 (53%) are female; of the patient group, 36 (43%) are female. Figure 3.1 (A) shows age distributions of the sample of recruits from IKEM. In the control group, a bias towards younger age can be observed. Some patients have available the *duration of untreated psychosis (DUP)* (period of time from first experienced psychotic episode to treatment onset) and/or the *duration of treatment (DT)* (at time of visit). Distributions of these durations are depicted in Figure 3.1 (D, E), showing that most patients have started treatment in the first three months after their first episode; most patients have also been undergoing treatment for three months or less at the time of the first visit, making them relatively pharmaco-naive. This stems from the inclusion criteria for ESO, which aim to recruit patients in their early phase of illness.

3.2 Correlations

Here we perform an exploratory correlation analysis of clinical features via Pearson's correlation coefficient for both IKEM and NUDZ datasets (Figure 3.2). In general, PANSS subscales are anti-correlated with age and correlated with each other, especially the positive and negative subscales with the global subscale (as also illustrated in Figure 3.1 (**F**). There is a weak positive correlation between DUP and visit age in both datasets and between DUP and positive symptoms in the NUDZ dataset.

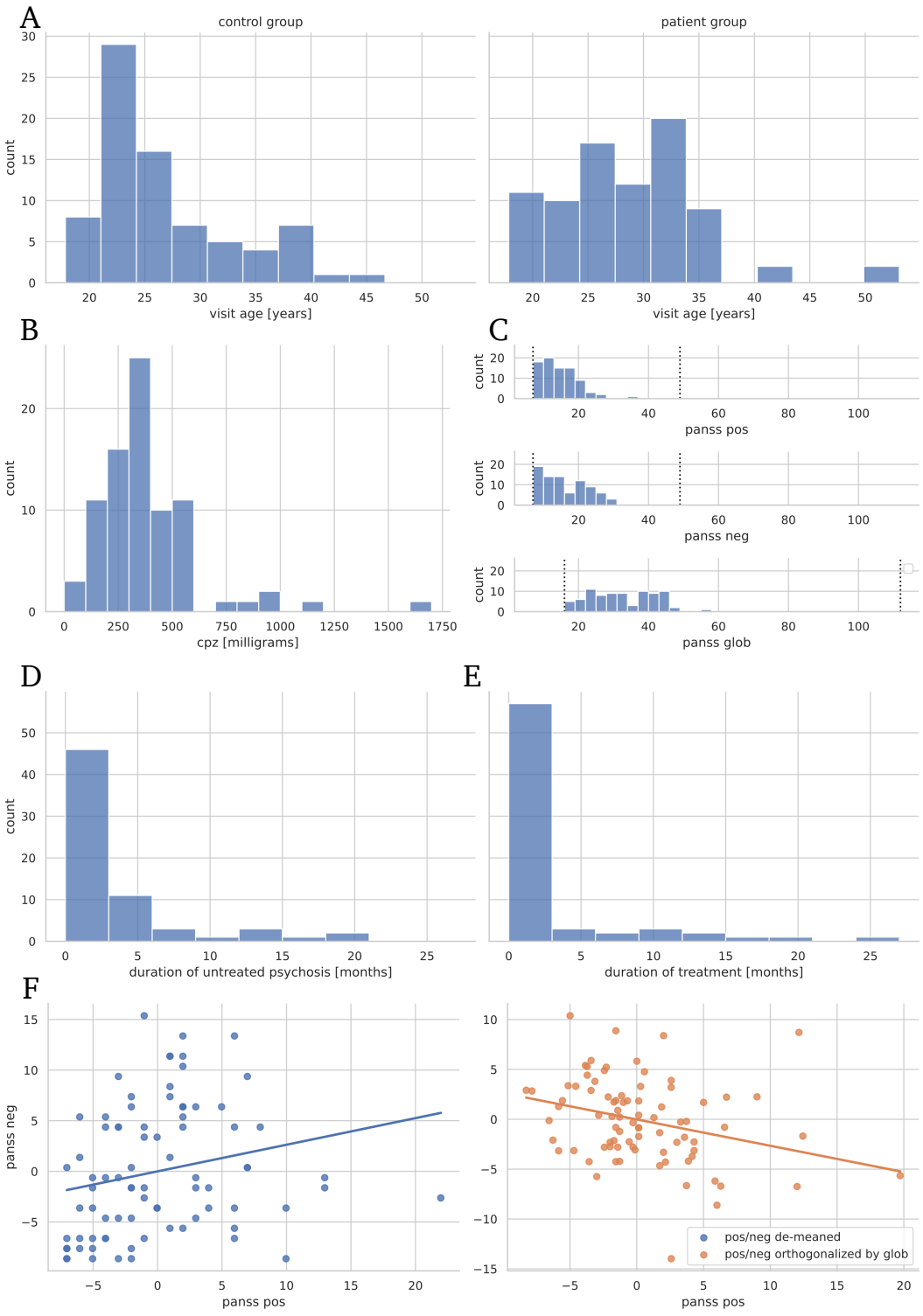

Figure 3.1: **A** Histograms: age at the time of visit for the control and patient groups (IKEM dataset). The mean and standard deviation of age is 26.9 (6.3) years in the healthy group and 28.6 (6.7) years in the patient group. **B** Histogram: chlorpromazine equivalent doses. **C** Histograms: PANSS subscales; subscale minima and maxima are delineated by vertical dotted lines. **D** Histogram: duration of untreated psychosis. **E** Histogram: duration of treatment. **F** Relationship between PANSS positive and negative subscales before and after orthogonalization by PANSS global.

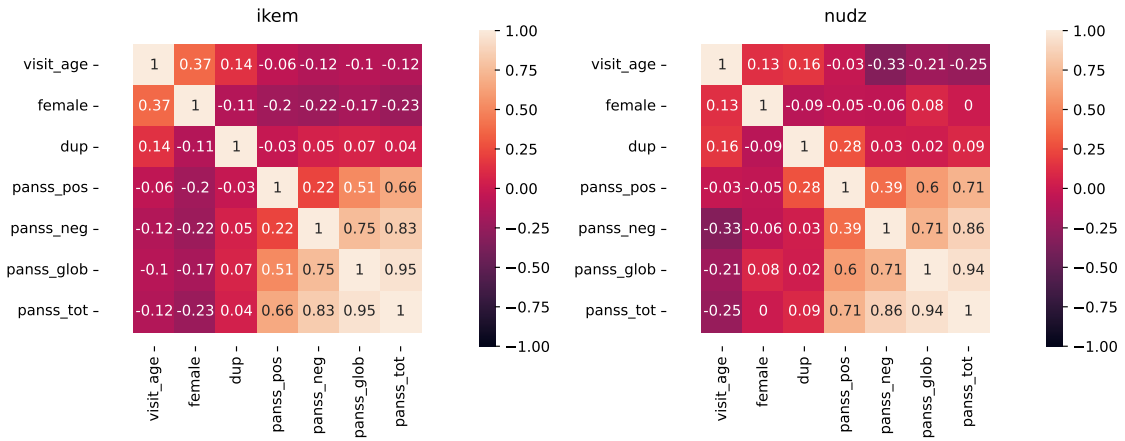

Figure 3.2: Correlation matrices for clinical variables.

3.3 Tools of the trade

In our analyses, we generally use this setup:

Orthogonalization We control for common demographic confounds by regressing out age and sex.

Standardization Especially in a multimodal setting, it is important to standardize features to zero mean and unit variance.

Principal component analysis We vary the number of components to probe the effect of dimensionality.

Linear regression We use L2 regularized linear regression with a constant penalty multiplier $C = 1$.

Logistic regression We use L2 regularized logistic regression with a constant penalty multiplier $\alpha = 1$.

Stratified 10-fold cross-validation (classification) Stratification of folds preserves class balance. We use 10 folds as it is the conventional number of folds balancing bias and variance of the estimates.

Repeated 5x2-fold cross-validation (regression) For regression we're not using strata, such as group membership or age ranges, so we use 2-fold cross-validation repeated 5 times on permuted samples.

3.4 Model overview

Although we fixed logistic regression for our subsequent analyses, we also carried out a survey of selected classical machine learning models. We did not grid search any hyperpa-

rameters and left the models in their "stock" settings. We understand that some predictors may not perform optimally with their default parameters (which also depends on the software framework in use, in our case scikit-learn); but we hope to see the comparative classification potential of the modalities rather than adapt the models to predict the outcome best, and also to see which predictors are generally well applicable in their default setting. The results are laid out in a cross-validated grid in Figure 3.3. FMRI modalities – FC and ALFF – show the largest classification potential, and linear models mostly outperform nonlinear models, which is expected at this sample size.

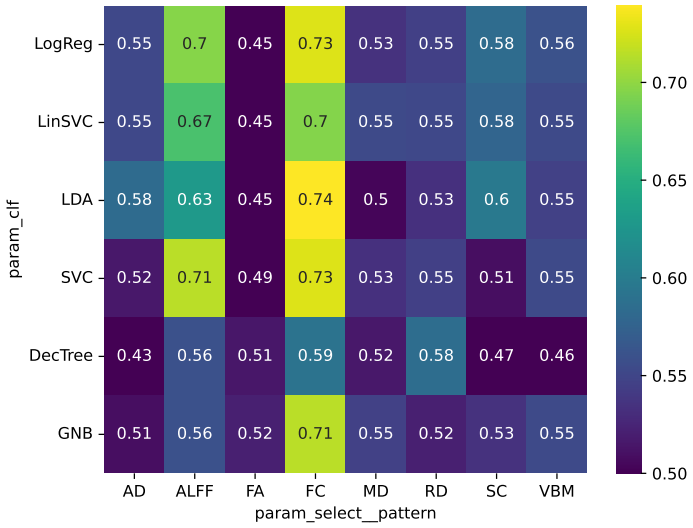

Figure 3.3: Mean accuracy for each modality-model combination. LogReg – logistic regression, LinSVC – linear support vector machines, LDA – linear discriminant analysis, SVC – support vector machines with radial basis functions kernel, DecTree – decision tree, GNB – Gaussian naive bayes

3.5 Unimodal components

We begin by grid-searching the PCA dimension for each modality separately to assess the relationship of modality and dimensionality. This means that for each number of components used to transform the data, a classifier was cross-validated. We searched only for up to 28 components because the smallest feature sets have 28 features, so we wanted to compare only up to this number. We will then choose the best model per modality for validation. Looking at Figure 3.4 **(A)**, both plots suggest that two feature sets – FC and amplitude of low-frequency fluctuations (ALFF) – dominate other modalities over most dimensions.

Table 3.1: Cross-validation results for top models per modality.

modality	mean accuracy	n components
ALFF	80.85%	9
FC	77.06%	10
VBM	63.42%	28
RD	61.47%	4
MD	59.60%	3
AD	57.13%	11
SC	55.88%	13
FA	55.26%	25

Findings 3.5

- Functional features lead in performance.
- VBM lags behind with weak accuracy.
- There is a stark difference between accuracy of the two connectivity metrics, FC and structural connectivity (SC).
- Diffusion features mostly perform about the same as random classification.
- Together with the fact that the diffusion features are missing in NUDZ, we will subsequently work only with FC, ALFF, VBM, and FA.

3.6 On the learning curve

In this analysis, we fix the predictor and its parameters and observe the scaling of predictive potential for each modality and across modalities. We use the same metamodel as in the previous section, with the number of PCA components set to 10. In Figure 3.5 we can observe the typical saturation behavior of learning curves on all of the modalities. As expected from the previous section, FC and ALFF lead by a large margin in test performance, whereas the performance SC does not seem to depend on the sample size, staying around 50% accuracy. Note that this behavior is recorded for 10 PCA components and may vary with dimensionality, but nevertheless this case illustrates learning rates for the modalities studied.

Findings 3.6

- FC and ALFF lead in learning curves, showing on par learning rate.
- Accuracy of SC does not improve with sample size.

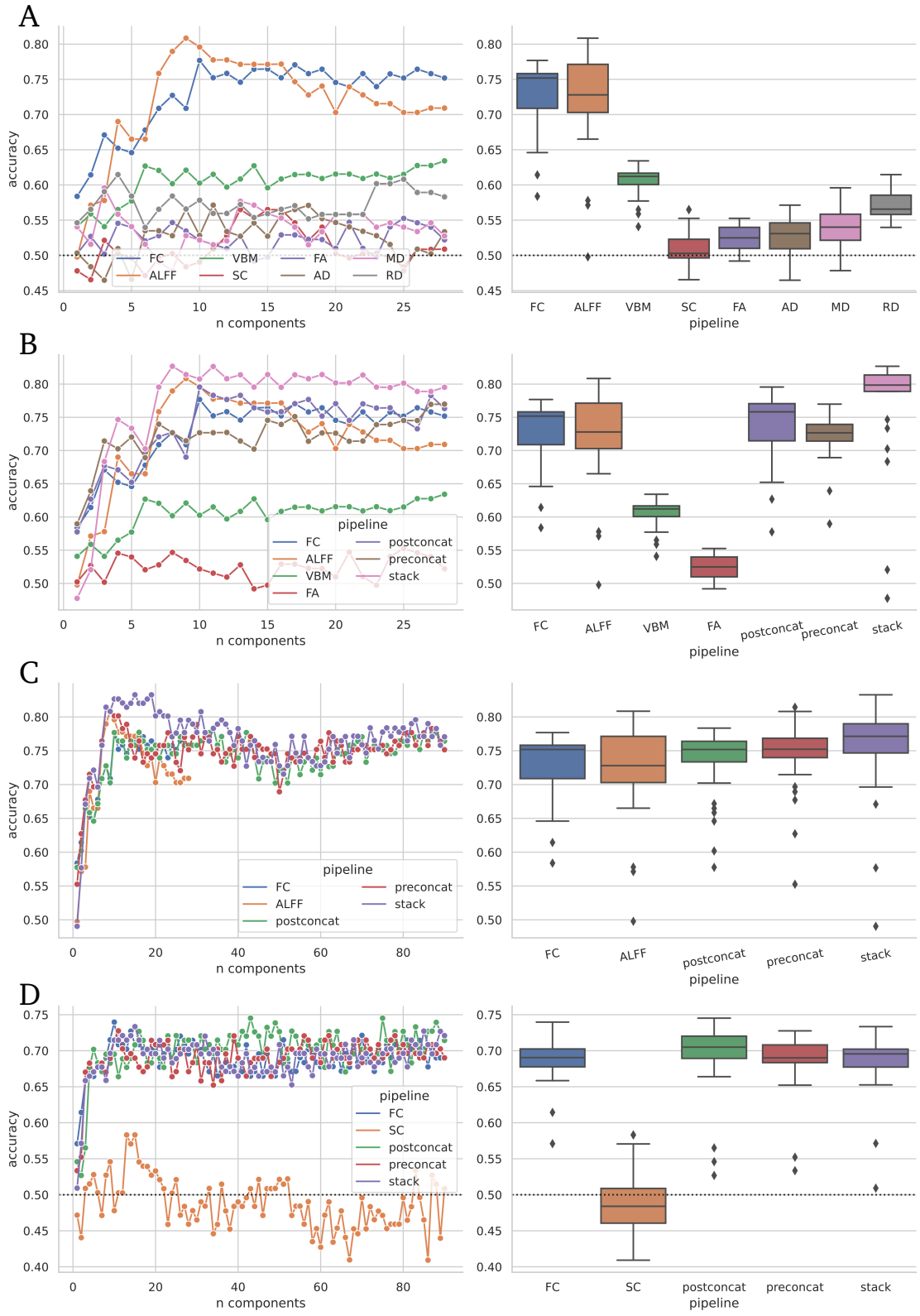

Figure 3.4: Lineplots showing accuracy versus number of components used in PCA and the boxplots showing the projections of all the results per modality. **A** For each modality separately. **B** For the three fusion strategies with all modalities as input, as well as with unimodal FC and ALFF plots for comparison. **C** For only FC and ALFF as input modalities to multimodal pipelines. **D** For FC and SC as input modalities to multimodal pipelines.

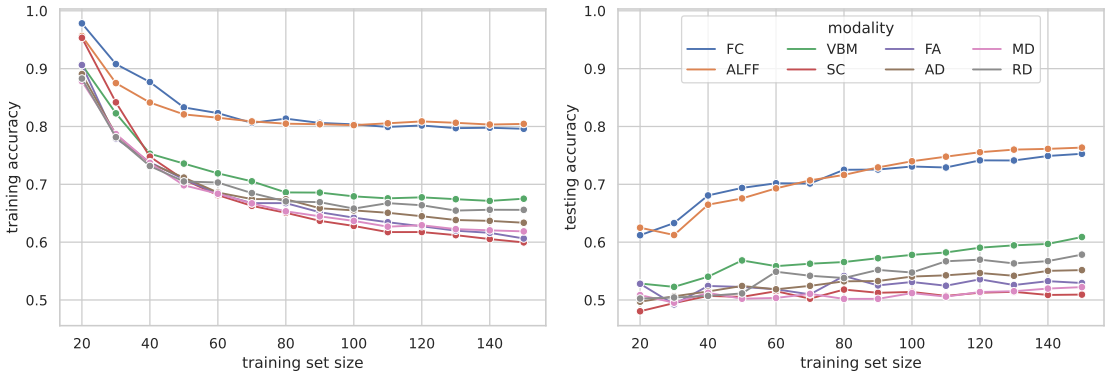

Figure 3.5: Learning curves for all IKEM modalities.

3.7 Multimodal fusions

We try several multimodal fusion strategies based on PCA.

postconcat – early fusion concatenates all feature sets together and applies PCA.

preconcat – early fusion applies PCA to each feature set separately and concatenates the components from all feature sets together.

stacking – late fusion applies PCA to each feature set followed by logistic regression, concatenating all the probability predictions. Then, a final logistic regression is performed (stacked) on top of these predictions to make the final prediction.

We again plot the performance on considered dimensions in Figure 3.4 (**B**), together with unimodal plots for FC and ALFF. The figure suggests that multimodal strategies actually do not improve accuracy over the best unimodal predictors, with the exception of the stacking classifier. We also have to note that for the preconcat and stacking the dimensionality is actually larger; each feature set yields n components, so the actual number of components used downstream in the predictor is multiplied by the number of input feature sets ($8 \times n$). This makes the curves as plotted for these pipelines incomparable with the unimodal ones, but they still serve as an indication of the dependence of pipeline performance on dimensionality.

Findings 3.7

- Stacking strategy visibly improved upon other pipelines.
- Blindly concatenating all possible features is not a straight way to improve accuracy – both concatenating strategies were outperformed by ALFF alone.

Table 3.2: Cross-validation results for top models per multimodal pipeline.

pipeline	mean test accuracy	n components
stack	81.0%	4×10
preconcat	77.7%	4×28
FC	75.8%	15
ALFF	74.0%	9
postconcat	73.9%	23
VBM	60.6%	6
FA	56.2%	4

3.8 Seek synergy

We have seen that FC and ALFF feature sets seem to dominate in performance. Here we explore the possibilities of the two feature sets working in synergy, separately from the other modalities. First, we perform an analogous grid-search like above, now inspecting components 1 through 90, as that is the number of features of the smaller feature set (ALFF). Figure 3.4 (**C**) illustrates a point: while both modalities perform moderately well, with ALFF performing better, the concatenated and reduced (postconcat) model performs on par with FC. We hypothesize that due to the sheer difference in dimensionality between the feature sets, FC features will have a major influence on the components yielded by PCA, thus yielding a classifier performing similarly to the unimodal FC one. We can observe multiple local maxima in the lineplot for ALFF classifiers that we would not see if searching only in the 1-28 component range. The preconcat and stacking method have yielded best results. On average, preconcat has achieved 78.3% accuracy at 11 components per modality (22 components total), and stack has achieved 77.7% at 29 components per modality (58 components total), as opposed to best ALFF accuracy of 76.4% at 68 components.

Table 3.3: Cross-validation results for top models using FC and ALFF.

pipeline	mean test accuracy	n components
preconcat	78.3%	2×11
stack	77.7%	2×29
ALFF	76.4%	68
postconcat	74.5%	50
FC	73.3%	10

Findings 3.8

- Multimodal methods combining FC and ALFF provide better accuracy than the unimodal models, with less overall components than the best ALFF model as well.
- It appears that taking away other modalities from the fusion strategies improves performance, when we focus on unimodally well performing feature sets.
- FC seems to degrade the performance when concatenated with ALFF before PCA dimensionality reduction.

3.9 Connectivity discrepancy

In this case, we explore the relationship between functional and structural connectivity. Both feature sets have the same large dimensionality, measuring connectivity between the 90 regions of the AAL atlas. However, as depicted in Figure 3.1 (**D**), there is a noticeable discrepancy between accuracy levels of FC and SC.

Table 3.4: Cross-validation results for multimodal pipelines using FC and SC.

pipeline	mean test accuracy	n components
postconcat	74.5%	43
FC	74.0%	10
stack	73.3%	15
preconcat	72.8%	11
SC	58.3%	15

Findings 3.9

- SC alone does not go above 60% accuracy, while FC achieved almost 75% on less PCA components.
- Any integration of SC with FC performs on par with unimodal FC.

3.10 Slippery (linear) slope

We have attempted to predict the PANSS scales of patients as a function of MRI modalities using principal component regression – PCA reduced features (5, 10, 15, and 20 components used) with a ridge linear regression model. Unfortunately, as is apparent in Figure 3.6, no model has achieved an R^2 larger than zero (as such, we did not further adjust

R^2). There is a systemic component to this phenomenon – R^2 is suitable as a goodness-of-fit measure on the training data, and when it is computed for predictions out-of-sample, the predictions of a model trained on the training set can be arbitrarily worse than those of a mean model of the testing set. When exploring unimodal prediction on individual components we see that ALFF or FC have in some cases achieved small positive R^2. This shows the weak point of regressing on the unsupervised PCA components, as some minor components with respect to maximizing variance within the regressor space can be more correlated with the target that, say, the first principal component is.

Findings 3.10

- Linear prediction does not perform well with respect to R^2, regardless of modality or target.
- Among the modalities, ALFF sometimes achieves small but positive R^2 when regressing on individual components Figure 3.6 (**C**).
- Among the targets, positive scale seems to have better potential for prediction than other scales, as highlighted in Figure 3.6 (**C**) when regressing on individual components.

3.11 Stratify patients

In the Section 3.10, we have seen some convincing arguments that linear prediction of clinical variables may not work well in practice, at least not in the naive way applied. To simplify the task, we have devised a simple strategy: using the median of each target variable (for all patients) as a threshold, and transforming the task into a binary classification. This will sort patients into groups of "lower" and "higher" PANSS scores. It is of note that this threshold is very much data dependent. A data independent threshold could be e.g. some quantile of the original scale range. We classified the patients grouped by the thresholded PANSS scores by the same unimodal pipeline we used in Section 3.5. The results are presented in Figure 3.7. The accuracy of classification is generally poor here compared to the classification into patients and controls, which is expected, as we are predicting within a single group on an arbitrary threshold. However, some models are still comparatively more applicable than the regression on untransformed PANSS (Figure 3.7 **A**. Even some of the univariate classifiers achieve above 60% accuracy. While more research would be needed, the stratification approach potentially could be of merit; however, it could be more data hungry if we wanted to to fill up the strata more uniformly or use a different threshold than the median, perhaps a data independent one. More than two strata also amplify the problem of small samples.

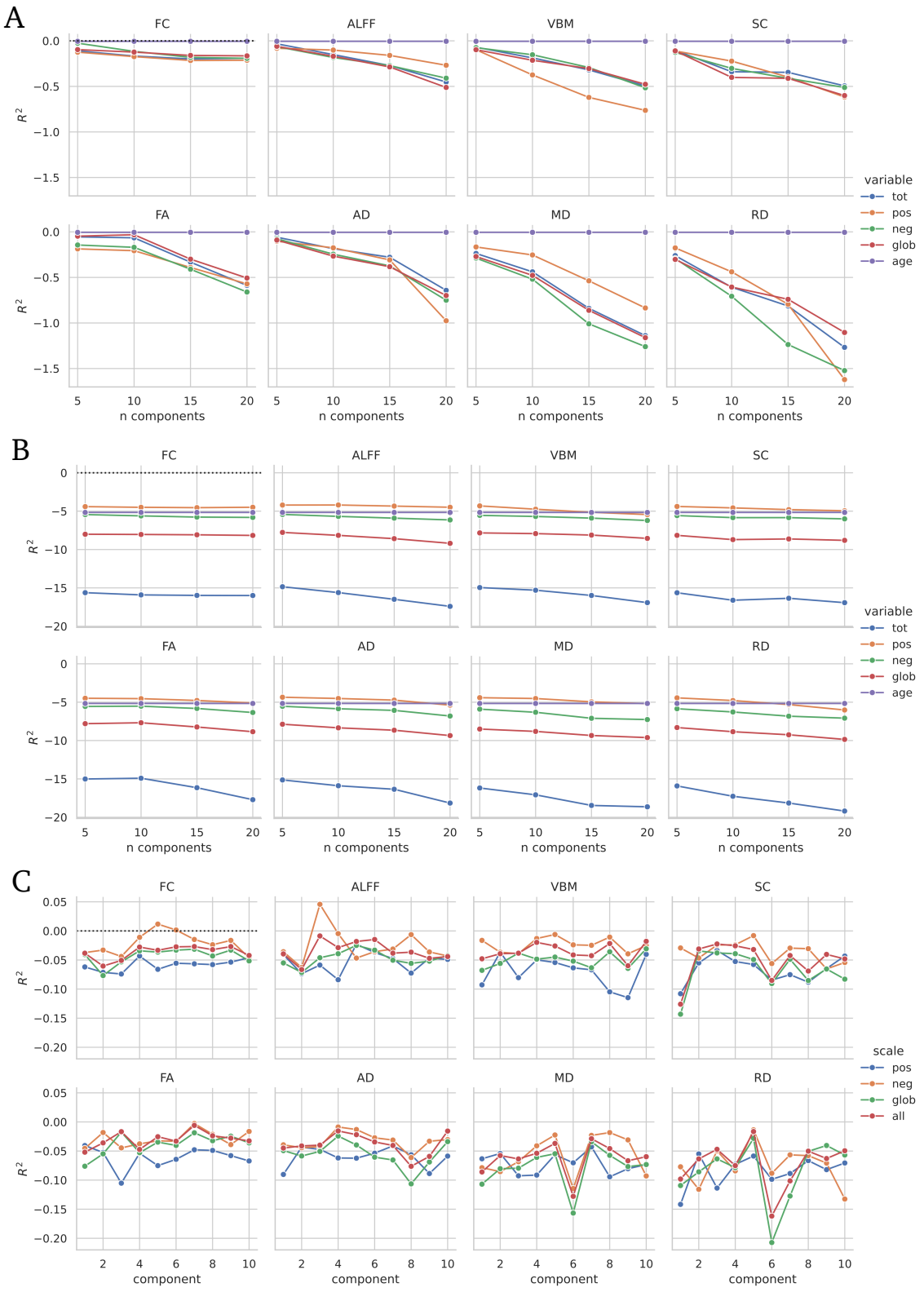

Figure 3.6: Linear regression evaluation with **A** R^2 and **B** mean absolute error. **C** Separate linear regressions were performed on each component individually to predict the PANSS scales, having better R^2 than the multivariate approach, yet still mostly negative.

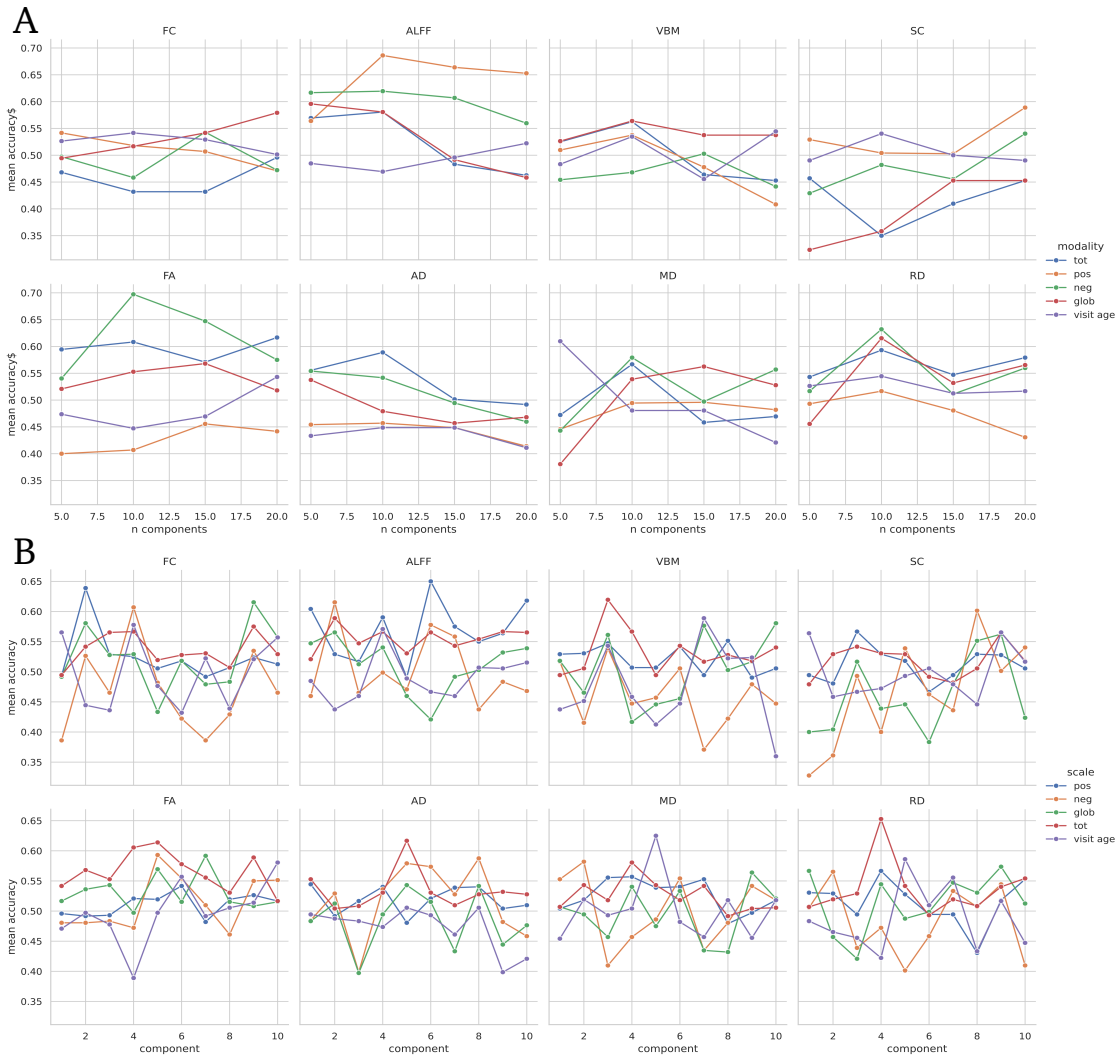

Figure 3.7: **A** Accuracy of classifying patients into above and below median variables. **B** Accuracy of classifying patients into above and below median variables on individual components.

Findings 3.11

- While most models in the stratification approach have poor performance, some ALFF and FA models achieve almost 70% accuracy.
- Some modality-scale combinations on individual components have peaks at around 60-65% accuracy.
- Using median and similar thresholds is data dependent, using fixed thresholds may bring up issues with class imbalance.

Chapter 4

Hypothesize & Validate

In exploring the IKEM dataset, we have made discoveries about the data and generated an array of questions. In this chapter, we will present these questions and answer them with results on the validation NUDZ dataset. We will work with models for four modalities: FC, ALFF, VBM, and FA. We selected models from these modalities by grid-search on IKEM, and then performed the same analyses on both datasets, comparing the results.

4.1 Which modality generalizes best?

We choose four models, one for each modality, with the number of PCA components that had the highest CV accuracy on the training (IKEM) dataset. All four models were cross-

Table 4.1: Training and validation accuracy for unimodal models.

modality	ALFF	FC	VBM	FA
n components	9	10	28	25
IKEM acc	80.85	77.68	63.42	55.26
NUDZ acc	62.81	72.50	60.40	55.93

Table 4.2: Training and validation accuracy for multimodal models.

pipeline	stack	postconcat	preconcat
n components	4×8	10	4×28
IKEM acc	82.7%	79.6%	77.0%
NUDZ acc	70.6%	68.2%	67.3%

validated in NUDZ dataset as well. Scores are summarized in Table 4.1. Cross-validated group predictions were used as an input to the Cochran's Q omnibus test to confirm differences in model performance. As a post-hoc test, the McNemar's test was used pairwise on the predicions of respective classifiers to determine pairwise differences. The omnibus

Figure 4.1: Assessment of parameter generalizability via accuracy comparison when selecting on IKEM (left) and NUDZ (right) and cross-validating on the other.

test found significant differences in performance both in the IKEM ($Q = 32.42$, $p < .001$) and NUDZ ($Q = 16.97$, $p < .001$) datasets. Seeing as the performance diminished when validating on the NUDZ dataset, we conducted the same procedure in the opposite direction – selecting on NUDZ and validating on IKEM – comparing the two in Figure 4.1. Here we can see that the least stable are FA and ALFF, with FC and VBM staying stable above 70%, and 60%, respectively.

Findings 4.1

- There are significant differences in performance between modalities in both datasets.
- On IKEM, ALFF significantly outperformed all other modalities except for FC.
- On NUDZ, the accuracy of all modalities decreased, with ALFF dropping most markedly, being significantly outperformed by FC.

4.2 Which multimodal pipeline generalizes best?

We have applied multimodal pipelines as defined in 3.7 to the tested modalities. Same comparisons as above were performed. Here, the pipelines fared similarly, and this was confirmed by Cochran's Q test, which did not detect any significant differences for either IKEM ($Q = 1.28$, $p = .52$) or NUDZ ($Q = 0.79$, $p = .67$). There was also a drop in accuracy from IKEM to NUDZ similarly to unimodal classifiers.

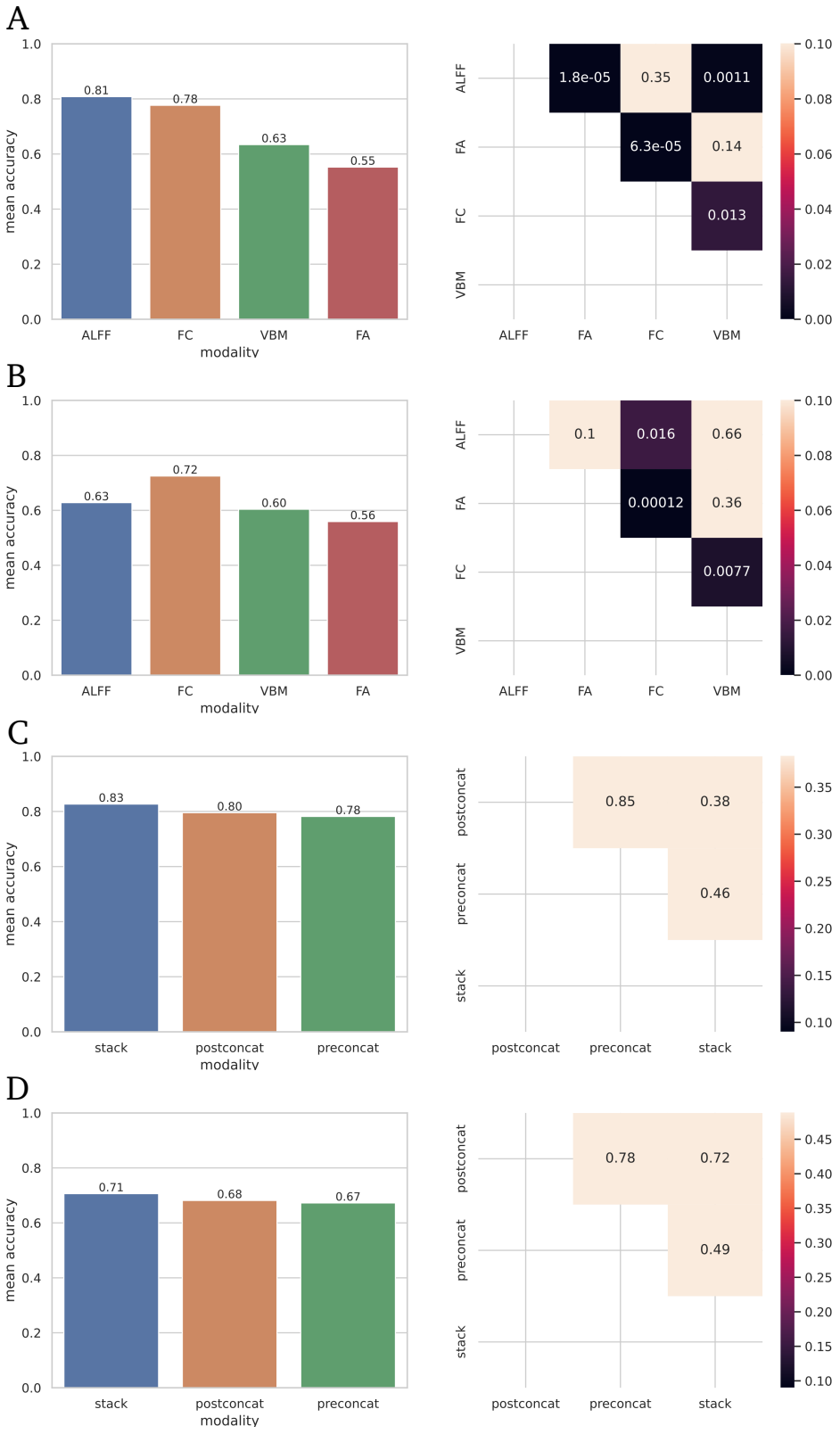

Figure 4.2: The cross-validated accuracy (left) and McNemar's pairwise test for performance comparison (right) for **A** unimodal classifiers (IKEM) **B** unimodal classifiers (NUDZ) **C** multimodal classifiers (IKEM) **D** multimodal classifiers (NUDZ).

Findings 4.2

- There are *no* significant differences in performance between fusion pipelines in both datasets.
- All three fusion pipelines had lower accuracy on NUDZ compared to IKEM.

4.3 Does reducing feature sets to FC+ALFF improve performance?

We use the multimodal pipelines applied to FC and ALFF with the best parameters found per pipeline in the exploratory phase. Cochran's Q does not find any significant difference between FC+ALFF pipelines in either dataset (IKEM: $Q = 3.71$, $p = .16$; NUDZ: $Q = 4.55$, $p = .10$). We then want to see the difference between the corresponding pipelines containing either 2 or 4 feature sets. For that, we again take predictions for each pipeline and compare with pairwise McNemar's test, but we do not find any significant difference.

Findings 4.3

- There is no significant difference between the pipelines using all four modalities and using FC and ALFF only.
- While the reduced model does not perform significantly better, it provides accuracy on par with the full model at lower dimensionality.

4.4 Does PANSS correlate with diagnosis probability?

Probabilistic predictions of classifiers studied above were obtained for all participants in the same way as for the class prediction for Cochran's/McNemar's test. Then, only patients were taken into account, and their individual PANSS subscores were correlated via Spearman's coefficient with their probability prediction. We hypothesize that the patients with larger PANSS scores should have higher probability of predicted diagnosis (or vice versa). There were some significant correlations in the unimodal IKEM classifiers, especially with FA. However, no correlations survive the Bonferroni correction, so the subscales do not correlate significantly with the PANSS scales in this framework.

Findings 4.4

- The probability of assignment to the patient group does *not* significantly correlate with the PANSS scores for both the unimodal and multimodal classifiers.

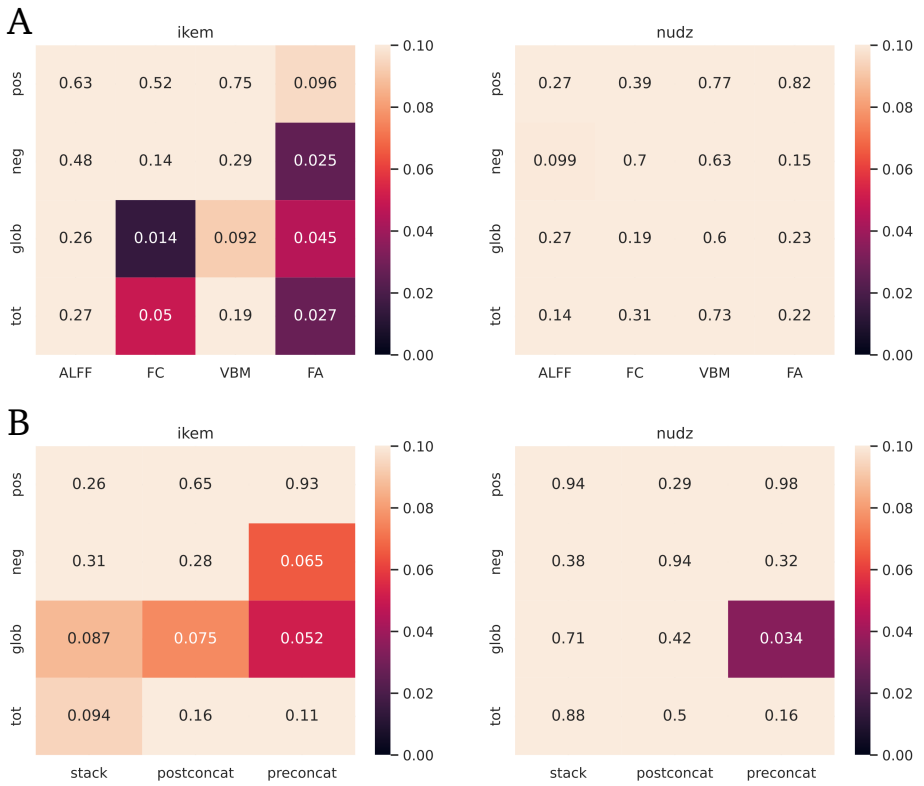

Figure 4.3: *P*-values of Spearman's correlations between PANSS scores and model predictions, for IKEM (left) and NUDZ (right) for **A** selected unimodal classifiers **B** selected multimodal classifiers.

4.5 How do learning rates compare across datasets?

We took the best unimodal models selected and used for analyses above and computed the learning curves for each model for both datasets.

Findings 4.5

- Learning curves show best learning rates for functional features, with ALFF falling to the level of VBM on NUDZ dataset.
- Structural connectivity slightly improves performance on training dataset, but not on validation.
- These results are in line with our analyses on the full sample size datasets.

4.6 Remarks

We have found that:

1. Of the validated unimodal models FC generalizes best.

2. Multimodal models generalize on par, with the stacking strategy being most promising.
3. Reducing the multimodal model to functional features performs about the same or better at reduced dimensionality.
4. Prediction probability is not Spearnman correlated with PANSS scores.
5. Learning curves show qualitative differences in predictor learning rates.

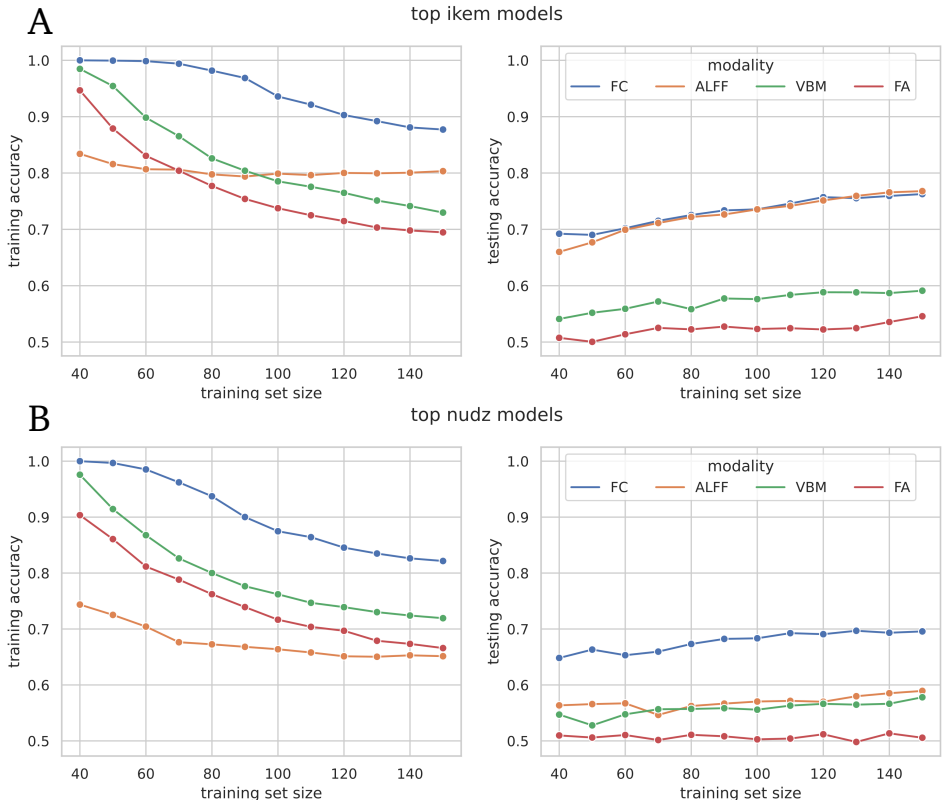

Figure 4.4: Learning curves for top unimodal models on the **A** IKEM dataset **B** NUDZ dataset.

Chapter 5

Ponder & Conclude

This thesis adds to the recently rapidly growing body of research on the integration of multiple modalities of MRI. We have utilized feature sets from three MRI modalities – structural, functional, and diffusion-weighted imaging – to assess the predictive ability in diagnosing schizophrenia from neuroimaging data. We've conducted a comparison between modalities in separation as well as designed several multimodal integration strategies for classification and regression analyses, and probed the effect of sample size on accuracy.

5.1 Towards individualized prediction

The driver of predicting illness is the optimization and better targeting of clinical care to improve the standard of living for people burdened with serious illness. The ability to detect schizophrenia would open the possibility of early intervention, even before the first symptoms or major episodes of the illness arise. However, the idea of simply trusting a model in clinical practice – even one with *reported* 99% accuracy – to label a *living* person with a diagnosis that could substantially alter their lives due to burdens of treatment seems infeasible as of yet. What seems to be achievable instead is to devise objective indicators, which would be useful together with wider evidence. There have been hopes for breakthrough medical indicators for a number of years (Lawrie et al. 2011; Iwabuchi, Liddle, and Palaniyappan 2013), however, to reach *clinical usefulness/utility*, one such marker has to be well tested and, more importantly, widely adopted. Another promising approach would be in the form of health recommender systems (Suphavilai, Bertrand, and Nagarajan 2018; Tran et al. 2021), suggesting appropriate treatment parameters, such as medication labels and dosage, possibly predicting expected drug side-effects, by comparing patient data. Studies such as ESO would be helpful in this aspect, relating functional outcome with neuroimaging and clinical data at several timepoints. All of that

said, we believe assessing the capabilities of unimodal and multimodal approaches to discriminate between healthy and patient groups is a first step towards individualized prediction in healthcare.

5.2 Potential of modalities

In this thesis, we have used preprocessed second-level features obtained by averaging voxel values across atlas regions. This approach can filter out a lot of regional noise and bring the feature distribution closer to the shape of the normal distribution, though it sacrifices the flexibility of early integration at the voxel level. As such, our results can only speak for model behavior at the regional level.

The structural modality, represented here by gray matter voxel-based morphometry features, had mediocre diagnostic power, reaching little over 60% accuracy consistently on both datasets. VBM accuracy in the range of 60-70% in the detection task is reported throughout the literature. While here we have limited ourselves to gray matter VBM, Salvador, Radua, et al. (2017) have conducted an assessment of a range of classifiers applied voxel-wise to both gray and white matter VBM and some other metrics such as wavelet-based morphometry. Importantly, the top accuracy of 75% in discriminating controls and patients was achieved with VBM as well (averaged over classifiers). This raises motivation to explore more in-depth the VBM data in the ESO dataset – first in trying to saturate the accuracy given the ROI based classifiers, and then possibly move to the voxel-based level. Another use of structural features could be found in anatomical stratification of individuals, such as when Dwyer et al. (2023) used previously found anatomical stratification of patients with psychosis and applied it to a mixed multi-site cohort of controls/patients, finding differences in, e.g., symptom remission at certain timepoints between patients in different subgroups. Similarly, Honnorat et al. (2019) found three neuroanatomical subgroups of patients. Findings like these motivate further inquiry into structural profile in psychosis as compared to the wider population, possibly then bringing back findings usable in individual prediction.

When it comes to diffusion-weighted MRI features, the jury is still out on whether they will be useful in schizophrenia analysis; their predictive performance in detecting schizophrenia has shown poor results so far, both in our analyses and the literature (Mikolas et al. 2018). In the case of structural connectivity, it is quite clear that as it stands, there is really no merit in it as a detector. Unimodally, it is no better than a coin-toss classifier, and there was no improvement when combined with functional connectivity as well. The SC as a modality is still quickly evolving with new tractography methods, as well as being quite dependent on the preprocessing steps, so we cannot forsake it as a whole, but we can confidently state that SC matrices were not effective as predictors in our study. Diffusion

tensor metrics have also fared poorly, mostly performing in the low end of the 50-60% range of accuracy.

These results are in accord with previous machine learning studies, whereas structural features generally show accuracy of around 60-70% (Yamamoto et al. 2020; Vieira et al. 2020; Winterburn et al. 2019). This is a stark contrast with their predictive power in detecting, e.g., multiple sclerosis (Rehák Bučková et al. 2022). However, detection is not the only task these modalities could be utilized in, and this utility may depend both on the way we preprocess and engineer these features and the target – these feature sets may perform better in predicting other clinical features, so we cannot rule them out in schizophrenia prediction yet. For instance, Tønnesen et al. (2020) uses diffusion tensor features for constructing brain-age models to assess deviations of schizophrenic and bipolar groups from the healthy control group in terms of predicted brain age. In our analyses, we have seen FA (along with ALFF) reach almost 70% accuracy when predicting the thresholded negative subscale of PANSS. These findings encourage finding ways of utilizing modalities even when working with an illness in which they do not play a primary role in detection.

For some time now, schizophrenia has been viewed through the lens of the disconnectivity hypothesis, referring to the improper integration of functional systems across the brain (K. J. Friston 1999), which suggests that functional features, specifically functional *connectivity*, should be the largest discriminators. Indeed, throughout our analysis, we have seen that functional features, FC and ALFF, perform substantially better than the anatomical or diffusion-based ones in detecting schizophrenia. This further adds to the findings reviewed in our survey of functional MRI prediction of schizophrenia in Section 1.8.2, with studies achieving accuracies often in the range of 80-90% and upwards using complex pipelines and nonlinear classifiers (Arbabshirani et al. 2013; Steardo et al. 2020). Kalmady et al. (2019) perform ensemble classification by incorporating several classifiers trained on a number of brain parcellations, and achieving high accuracy. This approach is interesting, as it takes into account the effect these different regional parcellations, which could be explored in research further. It is also of note that our linear models on functional features reached similar performance as the unimodal deep model by Rahaman et al. (2021).

Another avenue of research is generalizability. Fractional anistropy and voxel-based morphometry had weak accuracy similarly across both datasets, ALFF dropped from the best modality in IKEM to the level of VBM in NUDZ, while FC has retained its moderately high performance. Explaining the factors of this behavior may lead to a better understanding of the illness as well. Generalizability may be dataset-dependent, as different datasets may produce classifiers of lower or higher quality in this regard. Note that generalizability may be taken in different senses, such as model or meta-model/algorithm generalizability. The former would be when training the model on one dataset and testing it on the other. The latter is when taking the best-performing parameters on one dataset and fitting a sep-

arate model on the other dataset, assessing whether the parameters work as well across datasets, which is what we have done in our work.

5.3 Multimodal merit

We have seen that the multimodal models perform at least on par with the top unimodal model. This can be seen as a positive result, as it suggests the added information did not degrade the predictions. Importantly, when the unimodal logistic regressions were taken and a final logistic regression was stacked on top of the previous predictions (indeed a small neural network in principle), the performance seemed to improve. In general, the multimodal pipelines preserved fair amount of accuracy when validating the selected parameters on another dataset (see Figure 4.2 (**C**, **D**)). We have also seen in Section 3.8 and validation Section 4.3 that reducing and curating feature sets input into the learning algorithm may retain accuracy reached with a larger model or even improve it. These findings illustrate that before performing multimodal integration, one should be well acquainted with the modalities and features to be integrated in an attempt to maximize predictive potential. FC and ALFF features perform well enough on their own, often matching the multimodal pipelines in accuracy, so it is intuitive to try and find synergy between those two sets. The integration has shown no drop in accuracy from the model with all feature sets (reaching 84% CV accuracy in IKEM), which is a virtue in itself, as recording just a single imaging modality to achieve the same predictive power is more convenient. These results prompt a discussion about how to select and engineer the best features for prediction. In our review, we saw a parade of ways to use such approaches. Here, we performed an ad-hoc reduction of two second-level feature sets, as those generally performed well in our concrete analyses. Performing more detailed voxel-wise stepwise selection for the whole brain may still be quite infeasible due to computational constraints. One solution to this specific approach is outlined by Chin et al. (2018), where prediction is made voxel-wise, but the voxel features are selected in bulk as regions, so the number of iterations in the step-wise feature selection is drastically reduced.

5.4 Clinical correlations

Regression tasks on neuroimaging data in schizophrenia are not a very well-explored area compared to classification as of yet. As we have seen in our own analyses, there may be a good reason. In our analyses, regressing PANSS scores on brain data has not yielded optimistic results, with the goodness-of-fit measure R^2 showing largely negative values. Whether this is fault of using this measure in cross-validation, as it is designed for evaluation of goodnes-of-fit on training sets, is out of the scope of this thesis. The clinical scale

PANSS is not well for prediction. Most of the patients were positioned at the lower end of the scale (maybe intuitively, as they were early-stage schizophrenia patients on average), and the scale would probably need a large diversified sample to fill out the scale in the data. We have chosen to quantize the problem by thresholding each subscale by the median. This yielded interesting results, with some classifiers reaching about 70% accuracy, suggesting a suitably thresholded scale may be a promising target for classification, and motivating further in-depth exploration. One of the reasons we chose logistic regression is that it provides probabilistic output (without the need for Platt scaling and the like), which we deem a better alternative to hard classification, as it is a scale that could give a more gradual instrument for medical professionals. However, when we correlated patient PANSS scores with probabilistic predictions, we found no significant effect after correction, although looking at some p-values being below the significance level for FC and ALFF encourages further exploration of classifiers constructed with prior clinical information in mind. The regressions of clinical targets were done for patients only, however scales such as GAF or quality-of-life questionnaires are patient-agnostic and can be applied in constructing more comprehensive models for a wider population. A bigger merit of regressing of clinical scales and other variables might be in creating a sort of normative model rather than individual prediction, as demonstrated for age regression in the brain-age models by Tønnesen et al. (2020).

5.5 Size matters

Classification of schizophrenia patients from MRI already a wide area of research, with a number of unimodal (Arbabshirani et al. 2013; Kalmady et al. 2019; Salvador, Radua, et al. 2017; Vieira et al. 2020) and multimodal (Salvador, Canales-Rodríguez, et al. 2019; Ambrosen et al. 2020; Rahaman et al. 2021) studies reviewed in our work. The accuracy achieved in these studies varies greatly, often depending on confounding variables such as sample size (Arbabshirani et al. 2013), class imbalance, or multisite harmonization (Vieira et al. 2020). As we have stated earlier, we believe that a smaller sample size can unrealistically inflate performance metrics, such as in the fMRI study by Arbabshirani et al. (2013). Steardo et al. (2020, pp. 4–6) provide in Table 1 a well-structured overview of prediction studies using SVM on various modalities to predict schizophrenia, with several reaching up to 90% accuracy. It is of note that most of the studies have a modest sample size; in fact, most are smaller than our two sub-datasets taken separately: 73% are smaller than IKEM, median value of dataset:IKEM size ratio is 0.65; 82% are smaller than NUDZ, median value of dataset:NUDZ size ratio is 0.51. To support this claim with data, we collected the pairs (sample size, accuracy) from the table, taking the average where ranges were provided (visualized Figure 5.1), and computed Pearson's correlation coefficient,

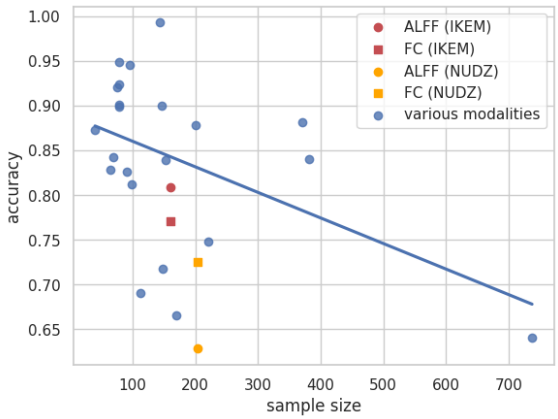

Figure 5.1: Accuracies and sample sizes from review by Steardo et al. (2020) with a linear trend and validation accuracies of functional features in our data.

which showed significant anti-correlation ($r = -0.47$, $p = .023$). While this is anecdotal evidence, it prompts further scrutiny of accuracy estimates with different combinations of modalities, predictors, and sample sizes, maybe constructing a sort of (approximate, due to study variability) "normative model of performance" per modality in the process. However, careful consideration should be in place, as, for example, the study with over 700 participants (Figure 5.1) is a multi-site study (6 sites), so it would be difficult to assess whether the weak accuracy is given by the sample size or the site-effect. The sample size may also play into the generalizability aspect of datasets.

5.6 Adverse site-effects

The effect of site of recording has been documented throughout literature, and there have been efforts in devising methods of multi-site harmonization (Johnson, Li, and Rabinovic 2007; Fortin et al. 2017; Bayer et al. 2022). While it is indeed important to see our trained models generalize across sites, the difference in scanner properties and local demography (Henrich, Heine, and Norenzayan 2010) make this approach much more complex than expected for MRI studies as of today (even more so for multimodal ones). The difference in the feature values caused plainly by site effect changes the interpretation from recognizing patients from controls to recognizing scanners/local populations. The demographic effect may be attenuated in ESO, because while the site changed during the project, the participant surveillance area has remained the same. Even then, as outlined in Table 2.3, there may be discrepancies in patient or sex ratios, as well as age distributions (although the latter two may be orthogonalized to some extent). However, the effect of the scanner still remains most apparent. Indeed, when we train a PCA model (including the standardization mean and standard deviation) on all IKEM features and project NUDZ features onto them, we get two well-separated distributions of data points (Figure 5.2). Although

both scanners in the ESO study have the same field strength, looking at the parameter descriptions in Section 2.2, we can notice differences in scanning sequence parameters, and even two versions of the fMRI protocol for the NUDZ machine. But it is not only a matter of scanner hardware. As is apparent from Figure 2.2, performance can vary substantially with preprocessing strictness. Even more sensitive to preprocessing and computation choices appears to be diffusion imaging, as can be illustrated when obtaining structural connectivity. When features are so heavily dependent on preliminary steps, it is key for researchers to open up raw data for research (after necessary anonymization). As matters stand today, it seems to be still too early to perform *reliable* cross-site prediction.

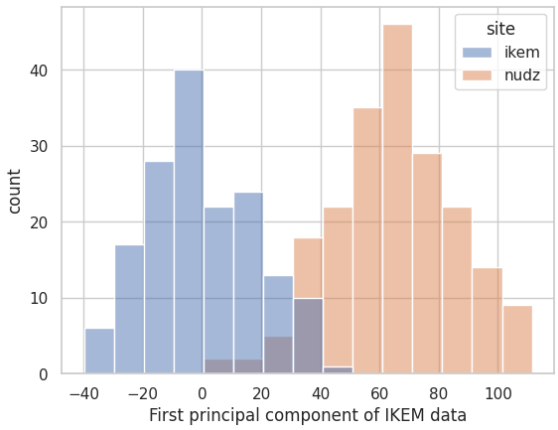

Figure 5.2: All features from IKEM and NUDZ projected to first principal component of IKEM.

5.7 Limitations of approach

We have scattered notes on limitations of studies and methods throughout discussion. Here we detail more specifically the limitations of our methodology. We do realize that this is a preliminary and exploratory study of potential data and methods, and as a consequence, our analyses are more qualitative in nature. As such, we take care not to make strong judgements on our concrete results. To minimize effects of methodology, we have fixed many of the parameters at some common, yet still indeed arbitary values, for example, the regularization strength C in our predictors. While controlling for these confounds may improve interpretability and lessen the complexity of our analyses, we must keep in mind that our outcomes are dependent on the values of such parameters. This principle applies not only to parameters, but also to conditions such as dataset permutation, i.e. the ordering of the training examples, which leads to different results when cross-validating the pipelines, as composition of individual folds matters. This effect can be mitigated to an extent by varying the number of folds, with leave-one-out CV, as it is not sensitive to permutations – given the number of folds is the same as number of data points, it exhausts

all the ways to split the dataset – but it is quite computationally expensive with larger samples and complex models, and there is a bias-variance trade-off in cross-validation as well. As such, we have conformed to one of the conventional numbers: 10 folds for classifier cross-validation, with stratification in folds provided by group labels. For regression we have used a different arrangement, the 5x2cv method, simply to see how this other popular method behaves on our data, and as an alternative to 10-fold CV with the same number of model fits. Also regarding cross-validation, we have used cross-validated predictions for use in the McNemar's and Cochran's tests, which is not usually found in literature, however, we haave justified this approach for the purposes of algorithm comparison. All of this is to say, we have fixed a number of variables, but the fixed values may be viewed as arbitrary or the justification may be lacking in completeness. We also draw reader's attention to the provenance of the data, which was obtained across two sites with inherent inconsistencies arising both due to the multi-site and long-running nature of the ESO project. As we have worked exclusively with the region-level extracted features, our judgements are constrained to this level of detail, so our assessment of modality potential might be quite skewed due to this fact.

5.8 Conclusions

In this thesis, we set out to perform and evaluate several machine learning approaches on the multimodal datasets from the ESO project. We have focused on second-level features derived from three imaging modalities: sMRI (VBM), fMRI (FC, ALFF), and dwMRI (SC, FA, AD, MD, and RD). We first constructed a pipeline that collects these feature sets from HYDRA data and merges them into two separate tabular datasets for data collected at IKEM and NUDZ. We then performed exploratory analyses on the IKEM dataset to get an overview of the properties of the data, as well as what can be achieved with the tools we have chosen to use. We then selected top-performing models that had the highest cross-validated accuracy and evaluated them on the NUDZ dataset as well. We couldn't use NUDZ as a testing set, as accounting for the scanner effect would be out of the scope of this work. As such, we compared the validated models within datasets and found out there are differences of performance between modalities. Functional connectivity performed similarly well across datasets, while ALFF, which by itself performed on par with multimodal pipelines on IKEM, had noticeably worse performance on NUDZ. Shaving down integrated modalities to just FC and ALFF has shown performance on par or higher than all four modalities, and though not significantly different, this suggests that curating features to more potent subsets could improve performance; conversely, that adding features without apriori information or feature engineering could introduce noise. On the other hand, integrating the functional and structural connectivity measures did not bring

any merit for classification. We have found that unimodal principal component regression does not predict PANSS scales well and recommend that different ways of transforming the predictors and targets should be explored while ideally preserving some interpretability of both. Another area of analysis of the patient group is stratification, which opens a wide realm of possible variables to stratify against. We have shown a clinical scale median-thresholding approach that yielded optimistic results, given the fact that we've classified within the patient group (then again, each subgroup was half the size of the patient group, which might have inflated the score). We have analyzed effect of sample size via learning curves, showing that the best discriminating features in our data may be reaching saturation at the size of the full sample, but similarly to the other analyses, we acknowledge that these behaviors are conditioned on our current data and experiment conditions. We have taken care to highlight these and other limitations throughout this work, whether in the review, the results, or the discussion. Finally, based on our research and results, we encourage further development of multimodal datasets and recommend deeper inquiry into modality-specific and multimodal applications in the quest of improving understanding of complex disorders such as schizophrenia.

Appendix A

Atlases

Table A.1: The 90 gray matter regions from the AAL atlas. Left hemispere indices are odd-numbered, right hemisphere indices are even-numbered.

Index	Left hemisphere	Index	Right hemisphere
1	Precental gyrus, left	2	Precental gyrus, right
3	Superior frontal gyrus, dorsolateral, left	4	Superior frontal gyrus, dorsolateral, right
5	Superior frontal gyrus, orbital part, left	6	Superior frontal gyrus, orbital part, right
7	Middle frontal gyrus, left	8	Middle frontal gyrus, right
9	Middle frontal gyrus, orbital part, left	10	Middle frontal gyrus, orbital part, right
11	Inferior frontal gyrus, opercular part, left	12	Inferior frontal gyrus, opercular part, right
13	Inferior frontal gyrus, triangular part, left	14	Inferior frontal gyrus, triangular part, right
15	Inferior frontal gyrus, orbital part, left	16	Inferior frontal gyrus, orbital part, right
17	Rolandic operculum, left	18	Rolandic operculum, right
19	Supplementary motor area, left	20	Supplementary motor area, right
21	Olfactory cortex, left	22	Olfactory cortex, right
23	Superior frontal gyrus, medial, left	24	Superior frontal gyrus, medial, right
25	Superior frontal gyrus, medial orbital, left	26	Superior frontal gyrus, medial orbital, right
27	Gyrus rectus, left	28	Gyrus rectus, right
29	Insula, left	30	Insula, right
31	Anterior cingulate and paracingulate gyri, left	32	Anterior cingulate and paracingulate gyri, right
33	Median cingulate and paracingulate gyri, left	34	Median cingulate and paracingulate gyri, right
35	Posterior cingulate gyrus, left	36	Posterior cingulate gyrus, right
37	Hippocampus, left	38	Hippocampus, right
39	Parahippocampal gyrus, left	40	Parahippocampal gyrus, right
41	Amygdala, left	42	Amygdala, right
43	Calcarine fissure and surrounding cortex, left	44	Calcarine fissure and surrounding cortex, right
45	Cuneus, left	46	Cuneus, right
47	Lingual gyrus, left	48	Lingual gyrus, right
49	Superior occipital gyrus, left	50	Superior occipital gyrus, right
51	Middle occipital gyrus, left	52	Middle occipital gyrus, right
53	Inferior occipital gyrus, left	54	Inferior occipital gyrus, right
55	Fusiform gyrus, left	56	Fusiform gyrus, right
57	Postcentral gyrus, left	58	Postcentral gyrus, right
59	Superior parietal gyrus, left	60	Superior parietal gyrus, right
61	Inferior parietal, but supramarginal and angular gyri, left	62	Inferior parietal, but supramarginal and angular gyri, right
63	Supramarginal gyrus, left	64	Supramarginal gyrus, right
65	Angular gyrus, left	66	Angular gyrus, right
67	Precuneus, left	68	Precuneus, right
69	Paracentral lobule, left	70	Paracentral lobule, right
71	Caudate nucleus, left	72	Caudate nucleus, right
73	Lenticular nucleus, putamen, left	74	Lenticular nucleus, putamen, right
75	Lenticular nucleus, pallidum, left	76	Lenticular nucleus, pallidum, right
77	Thalamus, left	78	Thalamus, right
79	Heschl gyrus, left	80	Heschl gyrus, right
81	Superior temporal gyrus, left	82	Superior temporal gyrus, right
83	Temporal pole: superior temporal gyrus, left	84	Temporal pole: superior temporal gyrus, right
85	Middle temporal gyrus, left	86	Middle temporal gyrus, right
87	Temporal pole: middle temporal gyrus, left	88	Temporal pole: middle temporal gyrus, right
89	Inferior temporal gyrus, left	90	Inferior temporal gyrus, right

Table A.2: The 50 white matter regions from the JHU atlas.

index	Area
1	Middle cerebellar peduncle
2	Pontine crossing tract (a part of MCP)
3	Genu of corpus callosum
4	Body of corpus callosum
5	Splenium of corpus callosum
6	Fornix (column and body of fornix)
7	Corticospinal tract, right
8	Corticospinal tract, left
9	Medial lemniscus, right
10	Medial lemniscus, left
11	Inferior cerebellar peduncle, right
12	Inferior cerebellar peduncle, left
13	Superior cerebellar peduncle, right
14	Superior cerebellar peduncle, left
15	Cerebral peduncle, right
16	Cerebral peduncle, left
17	Anterior limb of internal capsule, right
18	Anterior limb of internal capsule, left
19	Posterior limb of internal capsule, right
20	Posterior limb of internal capsule, left
21	Retrolenticular part of internal capsule, right
22	Retrolenticular part of internal capsule, left
23	Anterior corona radiata, right
24	Anterior corona radiata, left
25	Superior corona radiata, right
26	Superior corona radiata, left
27	Posterior corona radiata, right
28	Posterior corona radiata, left
29	Posterior thalamic radiation (include optic radiation), right
30	Posterior thalamic radiation (include optic radiation), left
31	Sagittal stratum (include inferior longitidinal fasciculus and inferior fronto-occipital fasciculus), right
32	Sagittal stratum (include inferior longitidinal fasciculus and inferior fronto-occipital fasciculus), left
33	External capsule, right
34	External capsule, left
35	Cingulum (cingulate gyrus), right
36	Cingulum (cingulate gyrus), left
37	Cingulum (hippocampus), right
38	Cingulum (hippocampus), left
39	Stria terminalis (can not be resolved with current resolution), right
40	Stria terminalis (can not be resolved with current resolution), left
41	Superior longitudinal fasciculus, right
42	Superior longitudinal fasciculus, left
43	Superior fronto-occipital fasciculus (could be a part of anterior internal capsule), right
44	Superior fronto-occipital fasciculus (could be a part of anterior internal capsule), left
45	Inferior fronto-occipital fasciculus, right
46	Inferior fronto-occipital fasciculus, left
47	Uncinate fasciculus, right
48	Uncinate fasciculus, left
49	Tapetum, right
50	Tapetum, left

Bibliography

1. Aas, IH Monrad (2010). "Global Assessment of Functioning (GAF): Properties and Frontier of Current Knowledge". In: *Annals of General Psychiatry* 9.1, p. 20. ISSN: 1744-859X. DOI: 10.1186/1744-859X-9-20.

2. Ambrosen, Karen S. et al. (2020). "A Machine-Learning Framework for Robust and Reliable Prediction of Short- and Long-Term Treatment Response in Initially Antipsychotic-Naïve Schizophrenia Patients Based on Multimodal Neuropsychiatric Data". In: *Translational Psychiatry* 10.1 (1), pp. 1–13. ISSN: 2158-3188. DOI: 10.1038/s41398-020-00962-8.

3. Arana, George W. (2000). "An Overview of Side Effects Caused by Typical Antipsychotics". In: *The Journal of Clinical Psychiatry* 61 (suppl 8), p. 13665. ISSN: 0160-6689.

4. Arbabshirani, Mohammad R. et al. (2013). "Classification of Schizophrenia Patients Based on Resting-State Functional Network Connectivity". In: *Frontiers in Neuroscience* 7. ISSN: 1662-453X. DOI: 10.3389/fnins.2013.00133.

5. Ashburner, John and Karl J. Friston (2000). "Voxel-Based Morphometry—The Methods". In: *NeuroImage* 11.6, pp. 805–821. ISSN: 1053-8119. DOI: 10.1006/nimg.2000.0582.

6. Azam, Muhammad Adeel et al. (2022). "A Review on Multimodal Medical Image Fusion: Compendious Analysis of Medical Modalities, Multimodal Databases, Fusion Techniques and Quality Metrics". In: *Computers in Biology and Medicine* 144, p. 105253. ISSN: 0010-4825. DOI: 10.1016/j.compbiomed.2022.105253.

7. Bach, Michael et al. (2014). "Methodological Considerations on Tract-Based Spatial Statistics (TBSS)". In: *NeuroImage* 100, pp. 358–369. ISSN: 10538119. DOI: 10.1016/j.neuroimage.2014.06.021.

8. Bayer, Johanna M. M. et al. (2022). “Site Effects How-to and When: An Overview of Retrospective Techniques to Accommodate Site Effects in Multi-Site Neuroimaging Analyses”. In: *Frontiers in Neurology* 13. ISSN: 1664-2295. DOI: 10.3389/fneur.2022.923988.

9. Berisha, Visar et al. (2021). “Digital Medicine and the Curse of Dimensionality”. In: *npj Digital Medicine* 4.1, pp. 1–8. ISSN: 2398-6352. DOI: 10.1038/s41746-021-00521-5.

10. Bitter, Istvan et al. (2008). “Antipsychotic Prescription Patterns in Outpatient Settings: 24-Month Results from the Intercontinental Schizophrenia Outpatient Health Outcomes (IC-SOHO) Study”. In: *European Neuropsychopharmacology* 18.3, pp. 170–180. ISSN: 0924-977X. DOI: 10.1016/j.euroneuro.2007.08.001.

11. Brown, Mark A. and Richard C. Semelka (2011). *MRI: Basic Principles and Applications*. John Wiley & Sons. ISBN: 978-0-470-92086-2. Google Books: oYOIHi3YkuMC.

12. Cabezas, Mariano et al. (2011). “A Review of Atlas-Based Segmentation for Magnetic Resonance Brain Images”. In: *Computer Methods and Programs in Biomedicine* 104.3, e158–e177. ISSN: 0169-2607. DOI: 10.1016/j.cmpb.2011.07.015.

13. Cao, H. et al. (2022). “Cerebello-Thalamo-Cortical Hyperconnectivity Classifies Patients and Predicts Long-Term Treatment Outcome in First-Episode Schizophrenia”. In: *Schizophrenia Bulletin* 48.2, pp. 505–513. ISSN: 0586-7614. DOI: 10.1093/schbul/sbab112.

14. Cao, Hengyi et al. (2018). “Cerebello-Thalamo-Cortical Hyperconnectivity as a State-Independent Functional Neural Signature for Psychosis Prediction and Characterization”. In: *Nature Communications* 9.1 (1), p. 3836. ISSN: 2041-1723. DOI: 10.1038/s41467-018-06350-7.

15. Chand, Ganesh B et al. (2020). “Two Distinct Neuroanatomical Subtypes of Schizophrenia Revealed Using Machine Learning”. In: *Brain* 143.3, pp. 1027–1038. ISSN: 0006-8950. DOI: 10.1093/brain/awaa025.

16. Chin, Rowena et al. (2018). “Recognition of Schizophrenia with Regularized Support Vector Machine and Sequential Region of Interest Selection Using Structural Magnetic Resonance Imaging”. In: *Scientific Reports* 8.1 (1), p. 13858. ISSN: 2045-2322. DOI: 10.1038/s41598-018-32290-9.

17. Cortes, Corinna et al. (1993). “Learning Curves: Asymptotic Values and Rate of Convergence”. In: Neural Information Processing Systems.

18. Dietterich, Thomas G. (1998). “Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms”. In: *Neural Computation* 10.7, pp. 1895–1923. ISSN: 0899-7667, 1530-888X. DOI: 10.1162/089976698300017197.

19. Dinga, Richard et al. (2021). *Normative Modeling of Neuroimaging Data Using Generalized Additive Models of Location Scale and Shape*. DOI: 10.1101/2021.06.14.448106. preprint.

20. *DSM-IV* (1994). Fourth edition. Washington, DC: American Psychiatric Association.

21. Dwyer, Dominic B. et al. (2023). “Psychosis Brain Subtypes Validated in First-Episode Cohorts and Related to Illness Remission: Results from the PHENOM Consortium”. In: *Molecular Psychiatry* 28.5, pp. 2008–2017. ISSN: 1476-5578. DOI: 10.1038/s41380-023-02069-0. pmid: 37147389.

22. Elad, Doron et al. (2021). “Improving the Predictive Potential of Diffusion MRI in Schizophrenia Using Normative Models-Towards Subject-Level Classification”. In: *Human Brain Mapping* 42.14, pp. 4658–4670. ISSN: 1097-0193. DOI: 10.1002/hbm.25574. pmid: 34322947.

23. Esteban, Oscar et al. (2019). “fMRIPrep: A Robust Preprocessing Pipeline for Functional MRI”. In: *Nature Methods* 16.1, pp. 111–116. ISSN: 1548-7105. DOI: 10.1038/s41592-018-0235-4.

24. Fortin, Jean-Philippe et al. (2017). “Harmonization of Multi-Site Diffusion Tensor Imaging Data”. In: *NeuroImage* 161, pp. 149–170. ISSN: 10538119. DOI: 10.1016/j.neuroimage.2017.08.047.

25. Friston, K. J. (1999). “Schizophrenia and the Disconnection Hypothesis”. In: *Acta Psychiatrica Scandinavica* 99.s395, pp. 68–79. ISSN: 1600-0447. DOI: 10.1111/j.1600-0447.1999.tb05985.x.

26. Fusar-Poli, P. and A. Meyer-Lindenberg (2016). “Forty Years of Structural Imaging in Psychosis: Promises and Truth”. In: *Acta Psychiatrica Scandinavica* 134.3, pp. 207–224. ISSN: 0001-690X, 1600-0447. DOI: 10.1111/acps.12619.

27. Geman, Stuart, Elie Bienenstock, and René Doursat (1992). “Neural Networks and the Bias/Variance Dilemma”. In: *Neural Computation* 4.1, pp. 1–58. ISSN: 0899-7667. DOI: 10.1162/neco.1992.4.1.1.

28. Girdhar, Rohit et al. (2023). *ImageBind: One Embedding Space To Bind Them All*. DOI: 10.48550/arXiv.2305.05665. arXiv: 2305.05665 [cs]. preprint.

29. Goto, Masami et al. (2022). “Advantages of Using Both Voxel- and Surface-based Morphometry in Cortical Morphology Analysis: A Review of Various Applications”. In: *Magnetic Resonance in Medical Sciences* 21.1, pp. 41–57. DOI: 10.2463/mrms.rev.2021-0096.

30. Hagmann, Patric et al. (2006). “Understanding Diffusion MR Imaging Techniques: From Scalar Diffusion-weighted Imaging to Diffusion Tensor Imaging and Beyond”. In: *RadioGraphics* 26 (suppl_1), S205–S223. ISSN: 0271-5333. DOI: 10.1148/rg.26si065510.

31. Henrich, Joseph, Steven J. Heine, and Ara Norenzayan (2010). “The Weirdest People in the World?” In: *The Behavioral and Brain Sciences* 33.2-3, 61–83, discussion 83–135. ISSN: 1469-1825. DOI: 10.1017/S0140525X0999152X. pmid: 20550733.

32. Hlinka, Jaroslav (2020). *[PROPOSAL] Predicting Functional Outcome in Schizophrenia from Multimodal Neuroimaging and Clinical Data*.

33. Hlinka, Jaroslav et al. (2024). “Role of fMRI Denoising for Classification of Schizophrenia from Functional Brain Connectivity”. In.

34. Honnorat, Nicolas et al. (2019). “Neuroanatomical Heterogeneity of Schizophrenia Revealed by Semi-Supervised Machine Learning Methods”. In: *Schizophrenia Research*. Machine Learning in Schizophrenia 214, pp. 43–50. ISSN: 0920-9964. DOI: 10.1016/j.schres.2017.12.008.

35. Huang, Shih-Cheng et al. (2020). “Fusion of Medical Imaging and Electronic Health Records Using Deep Learning: A Systematic Review and Implementation Guidelines”. In: *npj Digital Medicine* 3.1 (1), pp. 1–9. ISSN: 2398-6352. DOI: 10.1038/s41746-020-00341-z.

36. Iwabuchi, Sarina, Peter F. Liddle, and Lena Palaniyappan (2013). "Clinical Utility of Machine-Learning Approaches in Schizophrenia: Improving Diagnostic Confidence for Translational Neuroimaging". In: *Frontiers in Psychiatry* 4. ISSN: 1664-0640. DOI: 10.3389/fpsyt.2013.00095.

37. Jeurissen, Ben et al. (2019). "Diffusion MRI Fiber Tractography of the Brain". In: *NMR in Biomedicine* 32.4, e3785. ISSN: 1099-1492. DOI: 10.1002/nbm.3785.

38. Johnson, W. Evan, Cheng Li, and Ariel Rabinovic (2007). "Adjusting Batch Effects in Microarray Expression Data Using Empirical Bayes Methods". In: *Biostatistics* 8.1, pp. 118–127. ISSN: 1465-4644. DOI: 10.1093/biostatistics/kxj037.

39. Johnstone, Christopher Frith Eve (2003). *Schizophrenia: A Very Short Introduction*. Very Short Introductions. Oxford, New York: Oxford University Press. 216 pp. ISBN: 978-0-19-280221-7.

40. Jorge, João, Wietske Van Der Zwaag, and Patrícia Figueiredo (2014). "EEG–fMRI Integration for the Study of Human Brain Function". In: *NeuroImage* 102, pp. 24–34. ISSN: 10538119. DOI: 10.1016/j.neuroimage.2013.05.114.

41. Kalmady, Sunil Vasu et al. (2019). "Towards Artificial Intelligence in Mental Health by Improving Schizophrenia Prediction with Multiple Brain Parcellation Ensemble-Learning". In: *npj Schizophrenia* 5.1 (1), pp. 1–11. ISSN: 2334-265X. DOI: 10.1038/s41537-018-0070-8.

42. Kay, S. R., A. Fiszbein, and L. A. Opler (1987). "The Positive and Negative Syndrome Scale (PANSS) for Schizophrenia". In: *Schizophrenia Bulletin* 13.2, pp. 261–276. ISSN: 0586-7614. DOI: 10.1093/schbul/13.2.261. pmid: 3616518.

43. Kelly, Brendan D. (2005). "Structural Violence and Schizophrenia". In: *Social Science & Medicine* 61.3, pp. 721–730. ISSN: 02779536. DOI: 10.1016/j.socscimed.2004.12.020.

44. Lawrie Stephen M., STEPHEN M. et al. (2011). "Do We Have Any Solid Evidence of Clinical Utility about the Pathophysiology of Schizophrenia?" In: *World Psychiatry* 10.1, pp. 19–31. ISSN: 1723-8617. pmid: 21379347.

45. Le Bihan, Denis et al. (2001). “Diffusion Tensor Imaging: Concepts and Applications”. In: *Journal of Magnetic Resonance Imaging* 13.4, pp. 534–546. ISSN: 1522-2586. DOI: 10.1002/jmri.1076.

46. Lecun, Y. et al. (1998). “Gradient-Based Learning Applied to Document Recognition”. In: *Proceedings of the IEEE* 86.11, pp. 2278–2324. ISSN: 1558-2256. DOI: 10.1109/5.726791.

47. Leucht, Stefan et al. (2009). “Second-Generation versus First-Generation Antipsychotic Drugs for Schizophrenia: A Meta-Analysis”. In: *The Lancet* 373.9657, pp. 31–41. ISSN: 01406736. DOI: 10.1016/S0140-6736(08)61764-X.

48. Liu, Zhexing et al. (2010). “Quality Control of Diffusion Weighted Images”. In: *Proceedings of SPIE–the International Society for Optical Engineering* 7628, 76280J. ISSN: 0277-786X. DOI: 10.1117/12.844748. pmid: 24353379.

49. Manjón, José V. (2017). “MRI Preprocessing”. In: *Imaging Biomarkers: Development and Clinical Integration*. Ed. by Luis Martí-Bonmatí and Angel Alberich-Bayarri. Cham: Springer International Publishing, pp. 53–63. ISBN: 978-3-319-43504-6. DOI: 10.1007/978-3-319-43504-6_5.

50. Mikolas, Pavol et al. (2018). “Machine Learning Classification of First-Episode Schizophrenia Spectrum Disorders and Controls Using Whole Brain White Matter Fractional Anisotropy”. In: *BMC Psychiatry* 18.1, p. 97. ISSN: 1471-244X. DOI: 10.1186/s12888-018-1678-y.

51. Neal, Brady et al. (2019). *A Modern Take on the Bias-Variance Tradeoff in Neural Networks*. DOI: 10.48550/arXiv.1810.08591. arXiv: 1810.08591 [cs, stat]. preprint.

52. Oishi, Kenichi et al. (2008). “Human Brain White Matter Atlas: Identification and Assignment of Common Anatomical Structures in Superficial White Matter”. In: *NeuroImage* 43.3, pp. 447–457. ISSN: 10538119. DOI: 10.1016/j.neuroimage.2008.07.009.

53. Oldham, Stuart et al. (2020). “The Efficacy of Different Preprocessing Steps in Reducing Motion-Related Confounds in Diffusion MRI Connectomics”. In: *NeuroImage* 222, p. 117252. ISSN: 1053-8119. DOI: 10.1016/j.neuroimage.2020.117252.

54. Ooi, Leon Qi Rong et al. (2024). *MRI Economics: Balancing Sample Size and Scan Duration in Brain Wide Association Studies*. DOI: 10.1101/2024.02.16.580448. preprint.

55. Patel, Maxine X. et al. (2013). "How to Compare Doses of Different Antipsychotics: A Systematic Review of Methods". In: *Schizophrenia Research* 149.1-3, pp. 141–148. ISSN: 09209964. DOI: 10.1016/j.schres.2013.06.030.

56. Piorecka, Vaclava et al. (2022). "Extraction and Evaluation of EEG Covariates and Their Influence on GLM Model: EEG Covariates and Their Influence on GLM Model". In: *2021 International Symposium on Biomedical Engineering and Computational Biology*. BECB 2021. New York, NY, USA: Association for Computing Machinery, pp. 1–7. ISBN: 978-1-4503-8411-7. DOI: 10.1145/3502060.3502354.

57. Platt, John C (1999). "Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods". In.

58. Quirk, M E et al. (1989). "Anxiety in Patients Undergoing MR Imaging." In: *Radiology* 170.2, pp. 463–466. ISSN: 0033-8419. DOI: 10.1148/radiology.170.2.2911670.

59. Rahaman, Md Abdur et al. (2021). "Multi-Modal Deep Learning of Functional and Structural Neuroimaging and Genomic Data to Predict Mental Illness". In: *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. 2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC). Mexico: IEEE, pp. 3267–3272. ISBN: 978-1-72811-179-7. DOI: 10.1109/EMBC46164.2021.9630693.

60. Ramesh, Aditya et al. (2022). *Hierarchical Text-Conditional Image Generation with CLIP Latents*. arXiv: 2204.06125 [cs]. URL: http://arxiv.org/abs/2204.06125 (visited on 05/13/2024). preprint.

61. Rehák Bučková, Barbora et al. (2022). "Multimodal-Neuroimaging Machine-Learning Analysis of Motor Disability in Multiple Sclerosis". In: *Brain Imaging and Behavior*. ISSN: 1931-7557, 1931-7565. DOI: 10.1007/s11682-022-00737-3.

62. Rolls, Edmund T., Chu-Chung Huang, et al. (2020). "Automated Anatomical Labelling Atlas 3". In: *NeuroImage* 206, p. 116189. ISSN: 10538119. DOI: 10.1016/j.neuroimage.2019.116189.

63. Rolls, Edmund T., Marc Joliot, and Nathalie Tzourio-Mazoyer (2015). "Implementation of a New Parcellation of the Orbitofrontal Cortex in the Automated Anatomical Labeling Atlas". In: *NeuroImage* 122, pp. 1–5. ISSN: 1053-8119. DOI: 10.1016/j.neuroimage.2015.07.075.

64. Salvador, Raymond, Erick Canales-Rodríguez, et al. (2019). "Multimodal Integration of Brain Images for MRI-Based Diagnosis in Schizophrenia". In: *Frontiers in Neuroscience* 13. ISSN: 1662-453X.

65. Salvador, Raymond, Joaquim Radua, et al. (2017). "Evaluation of Machine Learning Algorithms and Structural Features for Optimal MRI-based Diagnostic Prediction in Psychosis". In: *PLOS ONE* 12.4, e0175683. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0175683.

66. Samartzis, Lampros et al. (2014). "White Matter Alterations in Early Stages of Schizophrenia: A Systematic Review of Diffusion Tensor Imaging Studies". In: *Journal of Neuroimaging* 24.2, pp. 101–110. ISSN: 1552-6569. DOI: 10.1111/j.1552-6569.2012.00779.x.

67. Schultz, Stephen H., Stephen W. North, and Cleveland G. Shields (2007). "Schizophrenia: A Review". In: *American Family Physician* 75.12, pp. 1821–1829.

68. Sheehan, D. V. et al. (1998). "The Mini-International Neuropsychiatric Interview (M.I.N.I.): The Development and Validation of a Structured Diagnostic Psychiatric Interview for DSM-IV and ICD-10". In: *The Journal of Clinical Psychiatry* 59 Suppl 20, 22–33, quiz 34–57. ISSN: 0160-6689. pmid: 9881538.

69. Škoch, Antonín et al. (2022). "Human Brain Structural Connectivity Matrices–Ready for Modelling". In: *Scientific Data* 9.1 (1), p. 486. ISSN: 2052-4463. DOI: 10.1038/s41597-022-01596-9.

70. Solmi, Marco et al. (2023). "Incidence, Prevalence, and Global Burden of Schizophrenia - Data, with Critical Appraisal, from the Global Burden of Disease (GBD) 2019". In: *Molecular Psychiatry* 28.12, pp. 5319–5327. ISSN: 1476-5578. DOI: 10.1038/s41380-023-02138-4.

71. *SPM12 Software - Statistical Parametric Mapping* (2024). URL: https://www.fil.ion.ucl.ac.uk/spm/software/spm12/ (visited on 05/20/2024).

72. Steardo, Luca et al. (2020). “Application of Support Vector Machine on fMRI Data as Biomarkers in Schizophrenia Diagnosis: A Systematic Review”. In: *Frontiers in Psychiatry* 11. ISSN: 1664-0640.

73. Suphavilai, Chayaporn, Denis Bertrand, and Niranjan Nagarajan (2018). “Predicting Cancer Drug Response Using a Recommender System”. In: *Bioinformatics* 34.22, pp. 3907–3914. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/bty452.

74. Tønnesen, Siren et al. (2020). “Brain Age Prediction Reveals Aberrant Brain White Matter in Schizophrenia and Bipolar Disorder: A Multisample Diffusion Tensor Imaging Study”. In: *Biological Psychiatry. Cognitive Neuroscience and Neuroimaging* 5.12, pp. 1095–1103. ISSN: 2451-9030. DOI: 10.1016/j.bpsc.2020.06.014. pmid: 32859549.

75. Tran, Thi Ngoc Trang et al. (2021). “Recommender Systems in the Healthcare Domain: State-of-the-Art and Research Issues”. In: *Journal of Intelligent Information Systems* 57.1, pp. 171–201. ISSN: 1573-7675. DOI: 10.1007/s10844-020-00633-6.

76. Tzourio-Mazoyer, N. et al. (2002). “Automated Anatomical Labeling of Activations in SPM Using a Macroscopic Anatomical Parcellation of the MNI MRI Single-Subject Brain”. In: *NeuroImage* 15.1, pp. 273–289. ISSN: 1053-8119. DOI: 10.1006/nimg.2001.0978.

77. Vieira, Sandra et al. (2020). “Using Machine Learning and Structural Neuroimaging to Detect First Episode Psychosis: Reconsidering the Evidence”. In: *Schizophrenia Bulletin* 46.1, pp. 17–26. ISSN: 0586-7614. DOI: 10.1093/schbul/sby189.

78. Wasserstein, Ronald L. and Nicole A. Lazar (2016). “The ASA Statement on P-Values: Context, Process, and Purpose”. In: *The American Statistician*. ISSN: 0003-1305.

79. WHO (2004). *The World Health Organization Quality of Life (WHOQOL) - BREF*. Geneva: World Health Organization.

80. Winterburn, Julie L. et al. (2019). “Can We Accurately Classify Schizophrenia Patients from Healthy Controls Using Magnetic Resonance Imaging and Machine Learning? A Multi-Method and Multi-Dataset Study”. In: *Schizophrenia Research* 214, pp. 3–10. ISSN: 09209964. DOI: 10.1016/j.schres.2017.11.038.

81. Yamamoto, Maeri et al. (2020). “Support Vector Machine-Based Classification of Schizophrenia Patients and Healthy Controls Using Structural Magnetic Resonance Imaging from Two Independent Sites”. In: *PLOS ONE* 15.11, e0239615. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0239615.

82. Yang, Hong et al. (2007). “Amplitude of Low Frequency Fluctuation within Visual Areas Revealed by Resting-State Functional MRI”. In: *NeuroImage* 36.1, pp. 144–152. ISSN: 1053-8119. DOI: 10.1016/j.neuroimage.2007.01.054. pmid: 17434757.