A multimodal modeling approach to schizophrenia

Axelsen, Martin Christian

Publication date:
2018

Document Version
Publisher's PDF, also known as Version of record

A multimodal modeling approach to schizophrenia

Martin Christian Axelsen

Kongens Lyngby 2018
PHD-2018-504

Technical University of Denmark
Department of Applied Mathematics and Computer Science
Richard Petersens Plads, building 324,
2800 Kongens Lyngby, Denmark
Phone +45 4525 3031
compute@compute.dtu.dk
www.compute.dtu.dk
PHD-2018-504

Summary (English)

Schizophrenia is a greatly invalidating disease with a largely unknown underlying etiology and pathophysiology. Attempts to investigate these difficult problems using a wide range of objective measures have been made over the past century. As the technological development has brought new advanced medical imaging modalities, combinations of views from multiple modalities have been investigated. The goal being to identify biomarkers for predicting the schizophrenia diagnosis, and furthermore to stratify schizophrenia patients into different subtypes.

In a time with focus of personalized medicine, single subject prediction has an increasing relevance. New statistical methods for analyzing the high dimensional multimodal data, and underdetermined problems are therefore needed.

In the present thesis, we have investigated two approaches to optimize the potential for using machine learning to analyze multimodal medical data. In a third study we have investigated the potential of using machine learning models to predict the diagnosis and prognosis of antipsychotic naïve first episode schizophrenia patients based on data from four different modalities.

In the fusion of multimodal data it is often unclear at what level or stage in an analysis pipeline the fusion will bring most information about a given problem. In the first methodological study of this thesis, we developed an approach to investigate the appropriate level of fusion of two modalities based on a hypothesis that the optimal fusion level is dictated by the dependencies among the modalities. The approach is based on two permutation schemes that establishes an upper and a lower bound for the integration level spectrum and attempts to

place the data set at hand on this spectrum.

In medical data, particularly multimodal data from intricate studies, data sizes are often modest. This can be a prohibitive factor for analyzing the data using machine learning models. With inspiration from comparative studies of learning efficiencies of the generative naïve Bayes model and the discriminative logistic regression model, and perspectives to data augmentation methods, we, in the second study, devise an approach to mitigate the small data problem. We suggest to generate synthetic data using the generative model, and use it to augment the training data of a discriminative model, with improved classification as a result in most cases.

In studying multimodal data from schizophrenia patients, knowing which modalities to rely on is nontrivial. In our study, we included four modalities of data from antipsychotic naïve first episode schizophrenia patients and healthy controls to investigate which modality had the best potential to predict the diagnosis of a subject using a range of machine learning models. Subsequently, the benefit of combining modalities and the potential to predict symptom remission of the patients were investigated. Our study showed that, against our expectation, only the cognitive modality had predictive capacities regardless of machine learning method. Multimodal combinations did not improve the performance, and the attempt to predict symptom remission in patients was not successful. However, a post publication study on applying the augmentation framework to potentially further improve the predictive capacity of the machine learning models on the cognitive modality looks promising.

Summary (Danish)

Skizofreni er en stærkt invaliderende sygdom med en stort set ukendt underliggende ætiologi og patofysiologi. Forsøg på at undersøge disse vanskelige problemer ved hjælp af en bred vifte af objektive mål er blevet udført igennem det sidste århundrede. Som følge af at den teknologiske udvikling har tilvejebragt nye avancerede medicinske billeddannelsesmodaliteter, er kombinationen af modaliteter blevet undersøgt. Målet værende at forsøge at identificere biomarkører til forudsigelse af skizofreni-diagnosen og endvidere for at stratificere skizofrenipatienter i forskellige subtyper.

I en tid med fokus på personlig medicin er individuel diagnoseforudsigelse af stigende relevans. Der er derfor et behov for nye statistiske metoder til analyse af højdimensional data og underbestemte multimodale problemer.

I den foreliggende afhandling har vi undersøgt to metoder til optimering af brugen af machine learning til analyse af multimodal medicinsk data. I et tredje studie har vi undersøgt potentialet for at forudsige diagnosen og prognosen af antipsykotisk naive første-episode skizofrenipatienter på baggrund af data fra fire modaliteter.

Ved fusion af multimodal data er det ofte uklart på hvilket niveau eller stadie af en analysefremgangsmåde, hvor fusionen vil bibringe mest information om et givent problem. I det første metodologiske studie i denne afhandling, udviklede vi en fremgangsmåde til at bestemme et passende fusionsniveau af to modaliteter baseret på en hypotese om at det optimale fusionsniveau er dikteret af afhængigheder imellem modaliteterne. Fremgangsmåden er baseret på to permutationsmetoder, der etablerer en øvre og en nedre grænse for integra-

tionsniveauspektrummet, og som forsøger at placerer det givne datasæt på dette spektrum.

Medicinske datasæt, især multimodale datasæt fra komplicerede studier, er ofte af beskedne størrelser. Dette kan forhindre en analyse af data ved hjælp af machine learning modeller. Med inspiration fra komparative studier af læringseffektiviteten af den generative naive Bayes model og den diskriminative logistiske regressionsmodel samt perspektiver til dataforøgelsesmetoder (data augmentation), udformer vi, i det andet studie, en tilgang til at afhjælpe problemet omkring små datasæt. Vi foreslår at generere syntetisk data, ved brug af den generative model, til at forøge mængden af træningsdata for den diskriminative model, med en forbedret klassifikation som resultat i de fleste tilfælde.

Ved studie af multimodal data fra skizofrenipatienter er det ikke trivielt, hvilke modaliteter man skal benytte. I vores studie inkluderede vi data fra fire modaliteter fra antipsykotika-naive første-episode patienter og raske kontroller for at undersøge, hvilken modalitet, der havde det højeste potentiale for at forudsige diagnosen hos et individ ved brug af en række machine learning modeller. Derefter blev fordelen ved at kombinere modaliteter samt muligheden for at forudsige symptomremission hos patienter undersøgt. Vores undersøgelse viste, at, i modsætning til vores forventninger, havde kun den kognitive modalitet kapaciteten til at forudsige diagnosen, uanset machine learning metode. Multimodale kombinationer forbedrede ikke ydeevnen og forsøget på at forudsige symptomremission hos patienter var ikke vellykket. I et studie udført efter udgivelsen af artiklen, undersøgte vi anvendelsen af den føromtalte dataforøgelsesprocedure til at forbedre forudsigelseskapaciteterne for machine learning metoderne på den kognitive modalitet. Resultaterne fra dette ser lovende ud.

Preface

This thesis was prepared partly at DTU Compute in the section of Cognitive Systems and partly at the Center for Neuropsychiatric Schizophrenia Research (CNSR) & Center for Clinical Intervention and Neuropsychiatric Schizophrenia Research (CINS), Mental Health Centre Glostrup, University of Copenhagen, Glostrup, Denmark in fulfillment of the requirements for acquiring a PhD degree in computer science.
The project was jointly financed by CINS and DTU.

The thesis deals with analysis of multimodal medical data from antipsychotic naïve first episode schizophrenia patients using machine learning methods.

The thesis consists of a summary report on relevant theory, synopses of the proposed methods and results from an analysis of clinical data, and perspectives of these. The thesis furthermore consists of three papers, of which one is published in a conference proceeding, one is in press and one is waiting to be submitted.

The thesis work was carried out in the period December 14, 2014 till December 28, 2018.

Kgs. Lyngby, 28-December-2018

Martin Christian Axelsen

Acknowledgements

First, I would like to thank my supervisors Nikolaj Bak and Lars Kai Hansen. I appreciate the many discussions of both scientific, academic and other natures.

I would also like to thank everyone from the sections at CNSR/CINS and at Cog-Sys for creating inspiring research environments and inspiring coffee breaks.

I am grateful for the Machine Learning and Neuroimaging Lab at University College London for hosting me in the fall of 2016. Particularly I would like to thank Janaina Mourãu-Miranda and João M. Monteiro for interesting discussions and the rest of the MLN lab and the Max Plank UCL Center for Computational Psychiatry and Aging Research for welcoming me.

I am thankful for the support I have received from the Otto Mønsted Foundation, Vera & Carl Johan Michaelsens Legat, the Augustinus Foundation and the Oticon Foundation for conference participation and the research stay at UCL.

Finally, I would like to thank my family for all their help. Sofie - you are the best! And Laurits, for giving me the push to get an early start of the day.

List of Contributions

Papers and manuscripts included in the thesis

Paper A: Axelsen, M. C., Bak, N., and Hansen, L. K. (2015). "Testing Multimodal Integration Hypotheses with Application to Schizophrenia Data." *In 2015 International Workshop on Pattern Recognition in NeuroImaging (PRNI)* (pp. 37-40). ieeexplore.ieee.org.

Paper B: Axelsen, M. C., Bak, N., and Hansen, L. K. 2019. "Transferability of learning efficiency from quick learning generative models." - To be submitted

Paper C: Ebdrup, B. H., Axelsen, M. C., Bak, N., Fagerlund, B., Oranje, B., Raghava, J. M., Nielsen, M.Ø., Rostrup, E., Hansen, L. K., Glenthøj, B. Y. (2018). "Accuracy of diagnostic classification algorithms using cognitive-, electrophysiological-, and neuroanatomical data in antipsychotic-naïve schizophrenia patients." *Psychological Medicine*, 1-10.

Published commentary not included in thesis

- Axelsen, M. C., Jepsen, J. R. M., and Bak, N. (2018). "The Choice of Prior in Bayesian Modeling of the Information Sampling Task." *Biological Psychiatry*, 83(12), e59-e60.

Contents

Chapter 1

Introduction

The common goal of this thesis is to investigate the disease of schizophrenia using the search lights of biological and statistical objective measurements. The profound biological variation in patients diagnosed with schizophrenia is however a great challenge, particularly in small data sets, which are common in the medical domain. The causal mechanisms for the biological variation are largely unknown, but have been found to be partly confounded by the duration of illness and exposure to antipsychotic medication [Fusar-Poli et al., 2013].

To approach these confounders, the focus of the present thesis is on antipsychotic naïve first episode (ANFE) schizophrenia patients, meaning that the patients have as short an illness duration as possible and have never been exposed to antipsychotic medication.

The data was collected at the Center for Neuropsychiatric Schizophrenia Research (CNSR) which is part of the Center for Clinical Intervention and Neuropsychiatric Schizophrenia Research (CINS) both primarily located at the Mental Health Center Glostrup, University of Copenhagen, Glostrup Denmark. At CNSR and CINS, ANFE schizophrenia patients have been recruited in three full cohorts (A-C), including baseline, short, and long term follow-up studies. A fourth ANFE schizophrenia cohort is currently in its recruitment phase. A range of objective measurements are collected in a range of modalities as listed in table 1.1. In this thesis, the data of focus is the cognition, electrophysiology, structural

Table 1.1: Overview of the CINS ANFE schizophrenia cohorts. In cognition, some tests are different across cohorts. In electrophysiology more paradigms are added in the later cohorts (B-E). For MRI measures cohort A was 1.5T, B-E 3T. Approximate numbers of patients (pt) and controls (ct) are shown as they vary across modalities. Cohort E is ongoing and numbers are hence aims.

pt / ct	**Cohort**	Psychopathology	Cognition	Electrophysiology	sMRI	DTI	^{1}H-MRS	fMRI	PET^{+}/SPECT*	GWAS
$\sim$ 25 / 25	**A**	x	x	PPI	x				^{123}I epidepride*	x
$\sim$ 45 / 45	**B**	x	x	PPI+	x	x		WM	^{18}F altanserin^{+}	x
$\sim$ 60 / 60	**C**	x	**x**	**PPI+**	**x**	**x**		Reward + RS	^{123}I iodobenzamide*	x
60 / 60	**E**	x	x	PPI+	x	x	x	Reward + RS	^{18}F DOPA^{+}	x

WM = Working memory paradigm, Reward = Reward paradigm, RS = Resting state,
PPI+ = pre-pulse inhibition (PPI), mismatch negativity (MMN),
P50 suppression, and selective attention (not included in paper C).
The modalities inside the box are included in paper C.

magnetic resonance imaging (sMRI) and diffusion tensor imaging (DTI) modalities from the baseline assessment of cohort C where all patients were ANFE schizophrenia patients. Across all CINS cohorts, additional modalities comprise: functional MRI (fMRI), ^{1}H magnetic resonance spectroscopy (MRS), positron emission tomography (PET) or single photon emission computed tomography (SPECT) scan, genome wide association study (GWAS) data and psychopathology (i.e. PANSS scores). Differences in research questions and development of new and improved techniques have caused differences in the recorded variables within the different modalities across the cohorts.

The contributions which are part of this thesis comprise

- A study investigating how to best fuse multimodal data (Paper A).
- A study on how to transfer the potentially superior learning efficiency of a generative to a discriminative model when working with small data sets (Paper B).
- A study on the diagnostic capabilities of machine learning methods applied to baseline data from ANFE schizophrenia patients and healthy controls (Paper C).

The rest of the thesis is structured as follows. In chapter 2 a brief introduction to the disease schizophrenia is given, general findings of the field using modalities similar to the ones analyzed in paper C are reviewed, and a short review of machine learning used to analyze such data in the field is given. In chapter 3 the focus is on multimodal medical data and how to analyze it using machine learning. Multimodal data fusion, general challenges related to this, and more specifically challenges related to the analysis of multimodal medical data are reviewed. In chapter 4, each of the three contributions included in this thesis are briefly reviewed and their main findings are highlighted. Finally in chapter 5, the findings of the three contributions are discussed and put into the mutual context of multimodal medical data from antipsychotic naïve first episode schizophrenia patients.

Chapter 2

Schizophrenia
- and how it is modeled

2.1 Brief overview of schizophrenia

Schizophrenia is a severe mental disease typically emerging in the early adulthood [Tandon et al., 2009]. It has a prevalence of approximately 0.5% and a world wide incidence of 7.7-43/100.000 (10-90 quantile) [McGrath et al., 2008]. The symptoms of the disease are often categorized into positive (e.g. delusions, hallucinations, and other reality distortions), negative (e.g. anhedonia, apathy, and social withdrawal), and cognitive deficits (e.g. memory, attention, and verbal fluency) [Tandon et al., 2009]. The etiology is widely unknown, though genetic factors and gene-environment interactions have been found to account for more than 80% of the liability for developing the disease [Tandon et al., 2008].

The diagnosis is generally established according to either the Diagnostic and Statistical Manual of mental Disorders which is currently in its fifth edition (DSM-5), published by the American Psychiatric Association [American Psychiatric Association, 2013], or according to the International Classification of Diseases, currently in its 11th edition (ICD-11) published by the World Health Organization (WHO) [World Health Organization, 2018]. Previous versions of the two guides have been found to be rather in agreement with each other

[Jakobsen et al., 2005, Fusar-Poli et al., 2016]. The diagnosis is solely established based on the presence of symptoms and impaired social function [Tandon et al., 2009] as no clinically useful biomarkers have been found so far [Prata et al., 2014].

The symptoms-based categorizations of mental diseases as implemented in ICD-11 and DSM-5 are, however, being criticized for not corresponding to treatment response [Cuthbert and Insel, 2013], hence indicating a heterogeneity in the disease. The Research Domain Criteria Initiative (RDoC) was introduced by the National Institute of Mental Health (NIMH) to encourage researchers to move away from symptoms-based investigations and focus more on e.g. treatment response [Insel, 2013].

Schizophrenia is generally regarded as a heterogeneous disease, which is illustrated by the difference in symptoms among patients and the difference in response to treatment. The heterogeneity of the disease is investigated from multiple perspectives, however often with one of two approaches: A subgrouping approach or a spectrum approach [Tandon et al., 2009, Kapur, 2011].

The subgrouping approach assumes multiple demarcated clusters of patients in a latent space. Patients within a cluster share common underlying biology and treatment response. The clusters thus represent different subtypes.

The spectrum, or dimension approach concerns the notion that the groups are not demarcated, but instead, the population is assumed to be situated on a continuum along a spectrum. If a person is situated sufficiently "far from the origin", the person will exhibit symptoms of the disease. There can hence exist multiple axes along which different strands of the spectrum can exist, corresponding to differences in biology, and response to treatment, i.e. different subtypes.

Several psychiatric diseases have been found to be heterogeneous (e.g. bipolar disease [Charney et al., 2017] and autism [Masi et al., 2017]). Furthermore, it has recently been found that there are shared genetic factors across a range of psychiatric diseases [Gandal et al., 2018] suggesting that these factors may explain an overlap of the diseases. Findings from several modalities are not specific to schizophrenia alone, but shared among several psychiatric diseases (see examples in [Gottesman and Gould, 2003, Goodkind et al., 2015]). This further suggests that the heterogeneity within psychiatric diseases reaches outside of the individual diseases.

As schizophrenia is a chronic disease, the treatment is focused on alleviating symptoms, but also on assisting with social aspects of coping with the disease e.g. via assisted housing and financial support. The treatment of symptoms

is primarily done using antipsychotic medication, however psychotherapy (e.g. cognitive behavioral therapy) and social treatment (e.g. supported employment) is also recommended [Tandon et al., 2010].

All existing antipsychotic drugs function as dopamine D2 antagonists, meaning that they block the dopamine D2 receptor. They have however primarily been found to be effective against the positive symptoms of the disease and less so on the negative symptoms and the cognitive deficits. Reductions in negative symptoms have been suggested as being secondary effects to the treatment of the positive symptoms [Tandon et al., 2010].

Based on the indirect evidence from the fact that all existing antipsychotic medication types are dopamine D2 antagonists and the fact that psychotic symptoms can be provoked through admission of dopamine enhancing drugs like amphetamine [Laruelle et al., 1996] make dopamine excess the oldest and most predominant hypothesis of the pathophysiology of schizophrenia. The dopamine hypothesis is however struggling to explain the non-positive symptoms [Howes et al., 2015].

Glutamate deficiencies constitute another major pathophysiological hypothesis of schizophrenia. It is primarily based on the N-methyl D-aspartate (NMDA) receptors since psychotic symptoms are found to be triggered by NMDA antagonists like phencyclidine (PCP) or ketamine. The glutamate hypothesis is in theory a better fit for explaining the broader schizophrenia symptom picture, including positive, negative and cognitive symptoms, however effects of treatment with glutamatergic drugs have not yet been successful. The imaging modalities used for investigating glutamate levels in-vivo also have caveats [Howes et al., 2015].

It is probable that neither hypothesis explains the full pathophysiology of all patients with schizophrenia, and that either a combination of both hypotheses, and likely more mechanisms involving e.g. neurodevelopmental factors and also other neurotransmitters compose a better explanation for the disease [Howes et al., 2015]. Another hypothesis is that each of the neurotransmitter hypotheses or a combination explains separate subtypes of the disease [Howes et al., 2015].

In investigating potential etiological factors of schizophrenia, doubt has been cast on the causality of biological differences found by empirical measurements. As an example, reduced brain sizes and increased ventricular volumes have been found rather consistently in patients suffering from schizophrenia, however these effects have also been linked to illness duration and the use of antipsychotic medication. This calls for a need for analyses of ANFE schizophrenia. Such a sample is the focus of the present thesis.

2.2 How to measure (major findings in the used modalities)

Statistically significant univariate group differences have been reported within schizophrenia for a range of different measures, however so far, none with robust diagnostic capabilities. Furthermore many findings are unspecific to schizophrenia, as they have been found to be related to other brain diseases as well.

Research in endophenotypes for schizophrenia has in recent years helped focus the search for more robust measures, which might also help explain the etiology of the disease. An endophenotype is broadly speaking a biomarker for a disease which has a genetic link, meaning that it is heritable and can thus also be found in unaffected relatives to a greater extent than in the general population [Gottesman and Gould, 2003]. An endophenotype must be associated with the disease in the population, however as a specific genotype can be associated with a range of phenotypes, an endophenotype is not necessarily disease specific.

In the following sections, findings within the selected modalities are highlighted. The modalities included in this section are the modalities which were included for analysis in paper C, and which comprise a subset of the data which is collected and investigated at CINS (see table 1.1). Results from the CINS baseline studies will be described at the end of the sections describing the respective methods.

2.2.1 Structural magnetic resonance imaging (sMRI)

MRI is a non-invasive radiological imaging technique. Strong magnets, magnetic field gradients, and radio waves are used in combination to excite hydrogen atoms in the water of the body, and subsequently measure the emitted response [Hanson, 2009]. MRI is often used for imaging the brain, i.e. neuroimaging, and includes a range of imaging methods, e.g. fMRI, DTI, and sMRI. sMRI is focused on imaging the anatomy of the brain, and particularly distinguishing gray matter (cortex) from white matter in the brain, and also to identify subcortical structures including the ventricles. The various parts of the brain can be segmented for further analysis using computational methods (e.g. FSL [Jenkinson et al., 2012]). For analyses of cortex, an atlas will often be applied to the imaged brain, to segment it into brain regions allowing identification of local differences. The process of applying an atlas includes warping the imaged brain to a standard space and subsequently applying the atlas (e.g. MNI standard space, and Harvard-Oxford cortical and subcortical structural atlases, as

in FSL). The choice of segmentation method and processing pipeline including atlas, relates to the desired output and hence relates to the hypothesis of the study.

It seems established in the field that reductions in whole brain and gray matter volumes, and increases in ventricular volumes exist when comparing schizophrenia patients with healthy controls [Keshavan et al., 2008]. More specifically, volume reductions seem to be driven by reduction in the gray matter volume, whereas the white matter volume seems to be unaffected [Haijma et al., 2013]. However, effects of age, lower IQ among patients with schizophrenia, smoking, drug abuse, unhealthy lifestyle, and the use of antipsychotics confound potential causal conclusions. Intracranial volume does however also seem to be reduced among patients, suggesting that the brain volume differences start early in life [Haijma et al., 2013]. Cortical differences have recently been analyzed in more detail in a meta-analysis by the ENIGMA group [van Erp et al., 2018]. They found that significant differences in cortical thickness and surface area exist when comparing patients with healthy controls. The difference in thickness seems to be region specific, and related to medication, whereas the difference found in cortical surface area is more global, and does not seem to be affected by the use of antipsychotics.

Generally, effect sizes of structural MRI differences are subtle [van Erp et al., 2018], and even more so when considering first episode patients [Haijma et al., 2013]

In the CINS cohorts no univariate volumetric differences have been found [Nørbak-Emig et al., 2017, Ebdrup et al., 2010]. In one cohort, cortical thickness, surface area, and curvature were investigated and significant group differences in curvatures of the left hemisphere were discovered [Jessen et al., 2018].

2.2.2 Diffusion tensor imaging (DTI)

DTI is used for imaging the structural connectivity of the brain by investigating diffusion of water molecules. As the intracellular diffusion is restricted by the cell anatomy of the axons, the diffusion provides an estimate of the architecture of the brain [Ambrosen, 2017]. Typical measures of DTI include fractional anisotropy (FA) and mean diffusivity, though mode of anisotropy, radial and parallel diffusivity can also be computed. In the recent cross-sectional meta-analysis by the ENIGMA Schizophrenia DTI Working Group [Kelly et al., 2018], widespread white matter changes were found across several brain regions of interest in schizophrenia. Lower global fractional anisotropy as well as widespread increases in mean and radial diffusivity were found for patients com-

pared to healthy controls. An initial effect of illness duration disappeared when controlling for age. No effect of medication was found. The authors are however calling for more longitudinal studies to investigate these factors properly.

DTI was only included in the later CINS cohorts (see table 1.1), where, using univariate analyses, subtle white matter integrity deficits in the patient group were found [Ebdrup et al., 2016].

2.2.3 Electrophysiology

When studying schizophrenia, common electrophysiological measurements comprise electromyographic (EMG) and electroencephalographic (EEG) measurements. EEG can be done in a resting state paradigm, but usually both EMG and EEG are used in paradigms where the response to a stimulus is recorded. In EEG, event-related potentials (ERPs) are investigated, where amplitudes of specific peaks following stimuli are compared. Electrophysiological measurements conducted as part of the CINS cohorts have varied across the cohorts, however they generally include pre-pulse inhibition (PPI), P50 suppression, mismatch negativity (MMN), and selective attention (though selective attention was not included in paper C).

In the PPI procedure a test subject is presented with a startle-eliciting stimulus where the response is measured using electromyographic measurements to quantify eye blinking. If the startle-eliciting stimulus is preceded by a pre-pulse in the form of a weak non-startling stimulus, the startle response is usually reduced. PPI deficiencies are seen in schizophrenia patients, however also in a range of other psychiatric disorders and diseases [Owens et al., 2016, Braff et al., 2001].

In the P50 suppression paradigm a pair of identical clicks of sound, typically 500ms apart, is presented to the subject. The P50 response is seen approximately 50ms after each stimulus, and in healthy controls, the response to the second stimulus will be decreased compared to the first. This is believed to be due to a sensory gating mechanism. The P50 suppression has been found to be reduced in patients suffering from schizophrenia, which is thought to be due to an impairment of the sensory gating mechanism. P50 suppression is generally regarded as a candidate for being an endophenotype for schizophrenia [Owens et al., 2016].

MMN is another ERP measure. The paradigm is setup as a sequence of regular frequent auditory or visual stimuli where deviant stimuli are interspersed. The MMN response is the difference between the EEG response to the deviant stimuli

and the frequent stimuli. In schizophrenia patients, the MMN has been found to be smaller than in healthy controls. MMN is also an endophenotype candidate [Owens et al., 2016], though it has not been found consistently in first episode schizophrenia patients [Haigh et al., 2017].

In the CINS cohorts, PPI deficiencies in patients were found to be significant in multiple cohorts [Mackeprang et al., 2002, Düring et al., 2014], though in one cohort this was only seen for males [Aggernaes et al., 2010]. P50 differences were found to be significant in one cohort [Oranje et al., 2013] but not another cohort [Düring et al., 2014]. MMN has not been found to be significantly different between patients and controls in any of the cohorts investigated [Oranje et al., 2017, Düring et al., 2015].

2.2.4 Cognition

Cognitive deficits are seen as symptoms of the disease, but are, as opposed to the positive and negative symptoms, objectively quantifiable. Nuechterlein and colleagues describe seven separate domains which they suggest for describing cognitive deficits in schizophrenia. They are: speed of processing, attention, vigilance, working memory, verbal learning and memory, visual learning and memory, reasoning and problem solving, and social cognition [Nuechterlein et al., 2004]. Deficits within domains can be evaluated using various test batteries, e.g. the Cambridge Neuropsychological Test Automated Battery (CANTAB) [Robbins et al., 1994].

Cognitive decline has been found to start before the onset of psychosis [Kahn and Keefe, 2013]. After onset of psychosis, reports of continued decline [Hedman et al., 2013] and reports of independence of duration of illness [Schaefer et al., 2013] indicate some uncertainty. Generalized cognitive impairment in schizophrenia has been established as a robust measure independent of cultural and linguistic differences and of changes in assessment criteria of the schizophrenia diagnosis over time [Schaefer et al., 2013]. Unfortunately, the deficits also seem to be relatively unaffected by the antipsychotic treatments that exist today [Kahn and Keefe, 2013].

In the CINS cohorts a range of cognitive deficits has been found that significantly differentiates patients from controls at baseline [Fagerlund et al., 2004, Andersen et al., 2011].

2.3 How to model using machine learning

Application of machine learning to schizophrenia data is wide spread (see e.g. [Orrù et al., 2012, Kambeitz et al., 2015, Wolfers et al., 2015, Arbabshirani et al., 2017]). Furthermore, in a recent paper, Bzdok and Meyer-Lindenberg seek to prime clinicians on the introduction of machine learning to psychiatry [Bzdok and Meyer-Lindenberg, 2018]. Despite the recent spectacular development in machine learning methods, no gold standard clinical diagnostic method with a reliable performance has been produced so far. One explanation could be related to the sizes of the data sets. Popular data sets used for benchmarking the progress in machine learning over the past decades include ImageNet with ∼3.2 million samples [Deng et al., 2009] and MNIST with ∼6000 samples [Lecun et al., 1998]). The data sets available within the field of schizophrenia, have observations on the order of tens or hundreds, and often have many variables. The problems are hence often underdetermined and the data too underpowered to allow training of high parameterized models like traditional deep learning models.

The primary focus, when applying machine learning analysis to schizophrenia data seems to be on neuroimaging techniques, especially MRI modalities (i.e. sMRI, fMRI, DTI) [Wolfers et al., 2015, Kambeitz et al., 2015, Orrù et al., 2012, Arbabshirani et al., 2017]. However, there is evidence that e.g. cognitive [Fagerlund et al., 2004, Andersen et al., 2011, Kahn and Keefe, 2013] and EEG [Owens et al., 2016] data also have potential of illuminating aspects of the disease. In paper C we investigated the accuracy of a range of algorithms using cognitive, electrophysiological and neuroanatomical (i.e. sMRI and DTI) data in ANFE schizophrenia patients.

Small data sets with high dimensionality, and the quest for clinical applicability have called for a search for biomarkers within psychiatric diseases in general, including schizophrenia [Orrù et al., 2012, Kambeitz et al., 2015]. Many studies are however still conducted as proof of concept studies [Kambeitz et al., 2015], where a search for variables as biomarker candidates is conducted on the same data set that is subsequently used for validation. Generalizability of the found variables or biomarkers is hence weakened and the predictive accuracies of the studies are artificially inflated [Arbabshirani et al., 2017].

Different machine learning methods have been applied to analyze schizophrenia data. Support vector machines (SVMs) are very popular within neuroimaging and brain disorders in general [Orrù et al., 2012, Kambeitz et al., 2015, Wolfers et al., 2015, Arbabshirani et al., 2017]. In a meta-analysis [Kambeitz et al., 2015] and a review [Wolfers et al., 2015] both on machine learning in psychiatry, support vector machines (SVMs) were used in about one third and half of the

included studies respectively. Another review is solely focused on the application of SVM in psychiatric diseases [Orrù et al., 2012]. Linear discriminant analysis (LDA) is also often used (∼30% in Kambeitz and 20% in Wolfers). The meta-analysis by Kambeitz and colleagues did not find significant differences in performances of SVM and LDA [Kambeitz et al., 2015]. However, they did see improvements using neural networks and deterioration when using random forest. Though, they could not conclude that the effects were not attributed to other differences between the studies.

Several steps are involved in analyzing schizophrenia data, starting at the raw acquired data to a final decision output of a machine learning classifier. The raw data, as acquired by a given modality is often high dimensional, particularly for the neuroimaging modalities, and also often noisy. Modality specific preprocessing is often used to condense the information of the modality. This can e.g. consist of calculating the relevant ERP measure from the raw EEG electrode signals (see e.g. [Oranje et al., 2013]) or computing specific domain scores from individual cognitive tests (see e.g. [Andersen et al., 2011]).

Following a modality specific preprocessing, a further aggregation of features may be applied. A reduction of the number of features can e.g. be obtained by feature selection using statistical methods. Another feature reducing technique is to apply unsupervised machine learning such as clustering or a form of matrix decomposition, such as the very popular principal component analysis (PCA) [Arbabshirani et al., 2017]. At this level in the analysis pipeline, imputation of missing data is often also done.

Finally a supervised machine learning method can be applied to learn the patterns in the data which potentially predicts a relevant output.

To ensure generalizability of the trained model, cross validation (CV) is often used in medical data. Usually, the modality specific preprocessing is done prior to entering the data in a CV framework, though the following steps of the processing pipeline should either be done using CV or trained on a completely separate data set.

Chapter 3

Multimodal medical data
- in a machine learning context

As earlier mentioned, strong univariate biomarkers ready for clinical implementation have not yet been found. A natural next step is therefore to look for multivariate biomarkers based on combinations of the modalities that individually show significant group differences. This calls for efficient methods for integrating data from different sources, known as data fusion.

Multimodal data has always been a part of the medical domain, however the data has often been fused via "visual inspection" [Calhoun and Sui, 2016] by the medical doctor. More specifically, the doctors have always considered multiple sources of information in order to make a decision. In recent years, the use of automatic multimodal analysis by data fusion has been rapidly increasing [Calhoun and Sui, 2016].

Data fusion, however, is in no way a trivial task. Choices must be made and challenges addressed, both of which potentially induce bias and restrict the possibility of performing ideal analyses.

In the following chapter an introduction to multimodal data fusion is given with a focus on analysis of medical data using machine learning methods. Choices and challenges are described, and unsolved problems that we have proposed solutions for in the contributions are illustrated.

3.1 Terminology

Multimodal data analysis and data fusion are relevant topics in a wide range of fields. Much of the basic research in data fusion stems from remote sensing and military applications, where e.g. satellite images and radar data are combined for investigating or tracking environmental development in an area or to track the position of a potentially hostile vehicle [Hall and Llinas, 1997, Pohl and Van Genderen, 1998]. Also multimedia research has contributed greatly to the field, where multimodal data in the form of audio, video and text is often combined to e.g. automatically infer the context of a movie clip [Atrey et al., 2010].

Within machine learning, data fusion has been covered under the term multi-view learning, which considers learning from multiple views, or data sets, to improve generalization [Zhao et al., 2017]. Multi-view learning methods are in the recent review by Zhao and colleagues [Zhao et al., 2017] separated in to three categories, where one category, termed co-regularization, includes the matrix decomposition method canonical correlation analysis (CCA) [Hotelling, 1936], which has been widely used to analyze multimodal medical data, particularly neuroimaging data [Calhoun and Sui, 2016].

The parallel development of methods for multimodal data fusion across fields has caused the terminology to develop in many directions, causing the terms "multimodality" and "data fusion" to be used under different synonyms and sometimes with other meanings. In the present thesis we will use the definitions as suggested by Lahat and colleagues:

Multimodal: "...a phenomenon or a system is observed using multiple instruments, measurement devices or acquisition techniques. In this case, each acquisition framework is denoted as a modality and is associated with one dataset. The whole setup, in which one has access to data obtained from multiple modalities, is known as multimodal." [Lahat et al., 2015]

Data fusion: "The analysis of several datasets such that different datasets can interact and inform each other." [Lahat et al., 2015]

3.2 Challenges related to multimodal data fusion

3.2.1 Data sizes

As mentioned previously, small data set sizes, i.e. data sets with data from relatively few subjects, are a general issue in medical data. It is both challenging to recruit patients and often also to find "matched" healthy controls to participate in a study. Obtaining a large cohort of participants for a study is particularly difficult when working with narrow inclusion criteria, such as first episode antipsychotic naïve schizophrenia patients as these specific requirements excludes many patients. The general findings of schizophrenia as a heterogeneous disease further indicate that many subjects are required in order to accurately describe the biological variation within the disease.

One way to mitigate the challenge of learning from small data set sizes is to refrain from using complicated models with a high number of parameters. In this domain it has been found that some models (e.g. naïve Bayes) require fewer observations in the training set to reach an asymptotic error compared to other models (e.g. logistic regression) even though the models are of comparable complexity [Ng and Jordan, 2002]. This encourages to primarily use the quicker learning models for medical data.

Another approach to the small data set challenge is to use data augmentation to artificially increase the size of the training set. This technique is widely used in deep learning (see e.g. [Krizhevsky et al., 2012]). Current data augmentation methods are, however, mostly developed for augmentation of image data sets via random rotations, cropping, and warping of the images [Hauberg et al., 2016]. These techniques do therefore not readily translate to non-image data. Recent attempts of data augmentation, that could work for non-image data, learn a feature space using a variant of sequence auto-encoders. In this latent space synthetic data is generated via interpolations and extrapolations between subjects of the same class [DeVries and Taylor, 2017]. It is thus similar to the well-known SMOTE algorithm for learning from imbalanced data sets [Chawla et al., 2002].

We investigated these two approaches to learning from small data sets in paper B.

3.2.2 Missing data

Missing data is another challenge when working with medical data which is also related to the difficulty of recruiting participants for a study. And the problem only grows when considering multimodal data. Despite the potentially more easily accessible healthy controls, missing variables are often seen for both patients and controls due to the elaborate and often expensive setup required to acquire data from the multiple modalities.
In this thesis, missing data will be categorized into four categories, all of which are likely to occur in multimodal data:

Missing completely at random:	(MCAR), typically when data gets lost or compromised e.g. a tube of blood is dropped on the floor and breaks.
Missing not at random:	(MNAR), when there is some relation between the missing variable and the variable itself, e.g. a patient is too ill to participate in the rest of the study.
Missing at random:	(MAR), when the missing variable is dependent on an observed value, e.g. the last measurements of a long battery of tests are more likely to be missing.
Block-wise missing	A phenomenon occurring in multimodal data, where a group of subjects may be missing data from one full modality, e.g. PET scans could be acquired from only a subset of subjects due to it invasive nature and high cost.

A simple and frequently employed way of handling missing data, is simply to exclude any subjects with missing data. This is termed complete case analysis. As the number of modalities grows, so will the probability that each subject is missing at least one variable. Complete case analysis will hence ultimately diminish the data set. Furthermore, it is argued that complete case analysis may introduce biases in the subsequent analyses [Donders et al., 2006].

Another way to handle missing data is to impute them, that is to represent them with an estimated numeric value instead of a missing value placeholder. Several imputation methods have been developed using various statistical approaches, though they often provide best estimates for MCAR and MAR [Donders et al.,

2006]. In so-called single imputation, a single complete data set is created by estimating the missing values, e.g. by using mean imputation, which inserts the mean value of the feature in place of the missing value. Particularly mean imputation has also been found to introduce biases [Donders et al., 2006]. In multiple imputation, several data sets are created hence attempting to create a distribution over the missing value, thereby providing a measure of uncertainty for the estimate. Multiple imputation has been found to reduce bias in the subsequent statistical analysis [Donders et al., 2006], however it can be computationally burdensome, particularly when implemented in a cross validation setting.

There are no established means of handling block-wise missing variables. Yuan and colleagues argue that imputation of the missing data is not feasible [Yuan et al., 2012]. They hence suggest to create smaller sub-data sets around the block-wise missing structures [Yuan et al., 2012, Xiang et al., 2013]. Attempts to impute block-wise missing data are however seen, using e.g. deep learning [Li et al., 2014] or principal component analysis [Zhu et al., 2018].

The issue of block-wise missing data can also be mitigated as part of the multimodal fusion strategy (see section 3.3.4).

3.2.3 Complementarity of modalities

The benefit of multimodal analysis over unimodal stems from the notion of diversity among the different modalities [Lahat et al., 2015]. Meaning that each modality contributes with a view of the problem that to some extent differs from the other modalities. Their views are then complementary to each other. The modalities can at one end of the spectrum be fairly similar in their views of the subjects. At the other end the modalities can contribute with completely independent views of the subjects, apart from all being able to identify something related to the target, i.e. they are conditionally independent, conditioned on the target variable.

Atrey and colleagues take up the discussion of which modalities to include with a focus on multimedia data [Atrey et al., 2010]. Modality selection is a more concrete task in multimedia analysis, where prior knowledge or more modality specific knowledge is available - e.g. when it is dark, a video feed provides less information. There are to the best of the present author's knowledge no existing modality selection schemes for multimodal medical data analysis. Modalities hence seem to be included on the basis of domain knowledge of unimodal or univariate group differences. An automatic approach to this is to fuse data according to the posterior probability of a well calibrated model based on unimodal

analyses as suggested in section 4.1.1. A non-informative modality would yield the same posterior probability for all classes, hence contributing with a "blank vote" for the fusion.

A modality which does not provide direct information about the target variable might, however, still contribute with a view of the noise in the data. Noise can be a relevant aspect of complementarity given that each modality contributes with a complementary view of the same sources of noise in the data. The fusion can then potentially assist in excluding these nuisances.

The degree of complementarity, the signal-to-noise ratio (SNR) of the complementarity, and of each modality affect how to best fuse the modalities. This is further accounted for in the following section.

3.3 Data fusion

Data from multiple modalities can be fused in different manners depending on the hypotheses for the analysis, the types of data, and restrictions in data, e.g. small data size or missing data.

In the following, an overview of different techniques for data fusion will be given, with a narrow focus on medical data.

3.3.1 Specific to image data

Generally within medical research, data often consists of images. This is also true for research within psychiatry, particularly computational psychiatry where neuroimaging is often used [Huys et al., 2016].

When fusing multiple neuroimaging modalities, an obvious benefit is the spatial aspect of the data. Spatial information from one modality can be used to guide the analysis of a second modality in a so-called asymmetric fusion [Calhoun and Sui, 2016]. This could e.g. consist of using the high spatial information obtained from fMRI scans of a subject to aid the EEG analysis of that same subject as done by Hansen and colleagues [Hansen et al., 2015]. Imaging data can also be fused symmetrically where each modality is analyzed concurrently [Calhoun and Sui, 2016]. In this setting, more general fusion methods which are independent of spatial or anatomical information can be used.

As image fusion is not within the scope of this thesis, this topic will not be discussed further.

3.3.2 Specific to temporal data

When operating with fusion of temporal data, e.g. in functional neuroimaging [Dähne et al., 2015], specific methods are needed to handle the different temporal resolutions across modalities. These methods are often referred to commonly as sensor fusion methods e.g. in autonomous driving, where particularly the Kalman filter seems to be a popular method of fusion [Khaleghi et al., 2013].

As static features were computed from the electrophysiology modality, which is the only temporal modality included in the work of this thesis, handling of temporal data is not within the present scope, and will hence not be discussed further.

3.3.3 How to fuse

When fusing modalities that do not necessarily have a spatial component, or when fusing image data with non-image data - e.g. fusing structural MRI data with data from various clinical tests [Monteiro et al., 2016], the spatial information cannot be utilized as a guidance for the analysis. Hence alternative approaches are needed.

The question of how to fuse can be answered in many ways, and the right answer is a compromise between several factors, which will be reviewed in section 3.3.4. In the present section, fusion levels and methods will be reviewed.

The fusion of multimodal data can generally be performed at different stages of the data processing pipeline as described in section 2.3, i.e. from very early in the process to very late. Again, different terminology for the fusion levels is being used in the literature. It is furthermore, also possible to combine different levels of fusion in hybrid fusions schemes where e.g. a few modalities are fused early and the resulting fused modality is then further fused with the remaining modalities at a later stage (see e.g. [Atrey et al., 2010]).

An overview of terminologies used in a selection of reviews from different fields is given in figure 3.1. The fusion methods are also often categorized into model-based methods and data driven methods. Model-based methods require a realistic model of the underlying process [Lahat et al., 2015] e.g. a model of the

biological pathogenesis for schizophrenia. Data driven methods learn the model directly from the data. Since the underlying mechanisms of schizophrenia are largely unknown, it is not feasible to construct a realistic model for schizophrenia. The focus of this thesis is therefore predominantly on data driven methods.

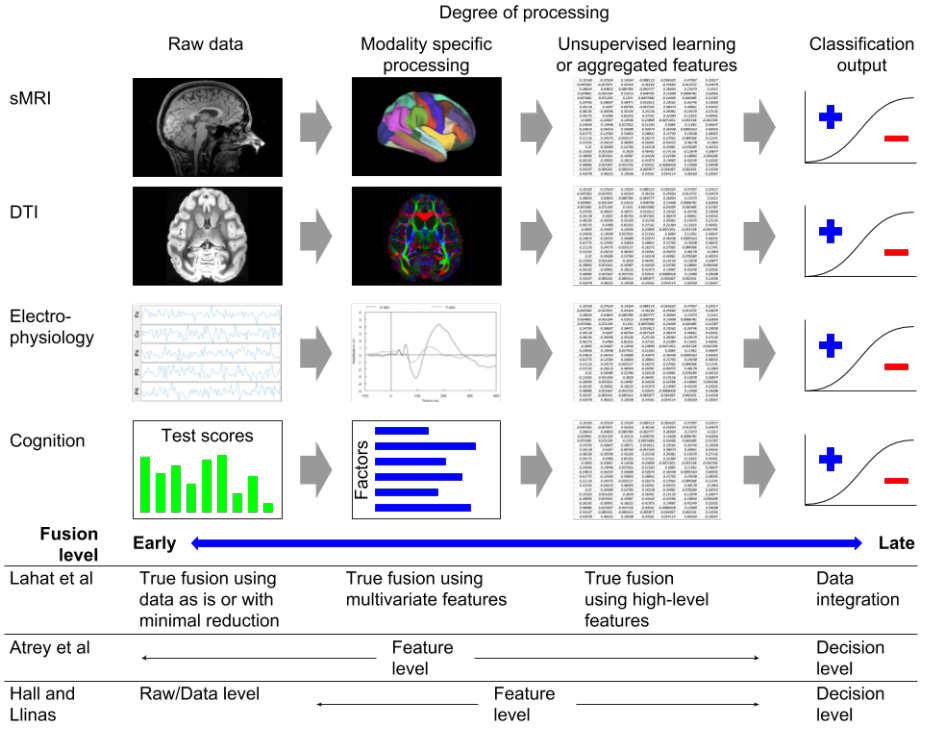

Figure 3.1: Conceptual illustration of the degree of processing of a range of modalities and the parallel fusion level ranging from early to late fusion. Below is an attempt to place the terminology from [Lahat et al., 2015], [Atrey et al., 2010], and [Hall and Llinas, 1997] on the spectrum. The sMRI image is of a patient from CINS cohort C [Ebdrup, 2009]. The illustration of the sMRI atlas (Desikan-Killiany atlas [Desikan et al., 2006]), and the DTI images are from [Ambrosen, 2017]. The EEG ERP is from [Oranje et al., 2013].

While in the following sections, the fusion level is separated into discrete categories, the actual level of fusion is more easily thought of as a spectrum spanning from early to late fusion.

3.3.3.1 Early fusion

In early fusion, the data is minimally processed prior to fusion. This entails that the data sets are in fact commensurable, i.e. that they are on the same scale and have the same spatial and temporal dimensionality, so that it is possible to fuse them. If the data sets are not commensurable, processing of the individual modalities is required prior to fusing. An earlier fusion should theoretically be ideal, as in this way, it will be possible to model dependencies between modalities directly [Lahat et al., 2015].

With a focus on classification, the commensurable variables from each modality can simply be concatenated to construct one large data matrix which can then be input to the subsequent analysis algorithm. The analysis can then be done with different priors to the individual modalities, or agnostically by treating all data equally. Alternatively a matrix decomposition method can be applied either on the concatenated matrix or in a fusion method such as CCA [Hotelling, 1936], where latent variables that are maximally correlated with all input modalities are identified. The goal of a matrix decomposition pipeline is often to identify one or more latent variables which in turn can be used for input to a machine learning classifier (or another statistical test). Matrix decomposition could ideally be trained on a separate data set in order to avoid overusing the small amount of data available. For a review on various matrix decomposition methods used in neuroimaging see [Sui et al., 2012].

3.3.3.2 Intermediate fusion

At an intermediate fusion level the data from each modality is more processed prior to fusing. A supervised feature selection method identifying informative features from each modality could for example be done prior to fusing. Supervised or unsupervised matrix decomposition methods could also be used on the modalities independently prior to fusion. This could e.g. be principal component analysis, which is often used in neuroimaging [Arbabshirani et al., 2017] possibly due to it being fairly robust against poor SNR values, particularly if adjusted for variance inflation [Abrahamsen and Hansen, 2011]. Often only a subset of the latent variables are included for further analysis [Hansen et al., 1999]. This requires the fusion to occur in a cross validation loop, or that the analysis is based on a training data set alone.

3.3.3.3 Late fusion

In late fusion the information from the separate modalities is fused at the decision stage of the processing pipeline. Often each modality is independently processed and analyzed using e.g. supervised classification, and the output of the classifiers from each of the modalities are then combined to form an ensembled decision about a test subject. The method for combining the outputs could be rule based [Atrey et al., 2010], similarly to what is typically done when machine learning classifiers are combined in ensembles [Kittler et al., 1998]. An example of this is majority voting, where each modality votes for an output class for a test subject, and whichever class receives most votes denotes the final output. The ensemble procedure could also be a weighted combination according to prior beliefs about the individual modalities or using a classifier to learn an optimal combination of modalities according to some true output. This again requires separate training and test data.

If the classifiers of the individual modalities provide posterior probabilities, the ensembling can also occur using Bayesian theory. We have theorized about this in Paper A. Another approach to handle fusion of probabilistic outputs is to use Dempster-Shafer (DS) theory. This is particularly relevant if the classes to be predicted are not necessarily mutually exclusive, as probability can be assigned to groups of classes as well as to individual classes. In DS theory, a lower bound called the belief and an upper bound called plausibility is determined for each class. A wide interval between the belief and the plausibility might suggest a need for more evidence, though this is disputed [Pearl, 1988]. A test subject can e.g. be assigned the class, or group of classes with the maximum belief [Xu et al., 1992].

3.3.4 Pragmatic and ideal approaches to data fusion

In the following section, choices regarding how to fuse are separated into two categories: Pragmatic choices and ideal choices.

Pragmatic choices in data fusion address the challenges related to multimodal data fusion as reviewed previously, and will typically move the level of fusion away from the very early stage. An example is when an early fusion of data yields a very underdetermined problem, the constraints needed to allow learning from this data set might be applying too great a bias. An alternative is hence to use a later fusion, where the dimensions can be reduced prior to fusing. This example illustrates the compromise needed, as none of these solutions necessarily provide unbiased results. Another issue is related to missing data, particularly block-

wise missing data, which is not readily imputable as reviewed in a previous section. As opposed to imputing the data, a late integration pipeline where each modality is modeled independently can be build up around the block-wise missing data. Each model would here only be trained on the data that is available for the given modality and only contribute in the decision stage to the test subjects where the given modality is actually available.

The pragmatic reasons for fusion are generally related to the signal-to-noise ratio in the data, where if it is sufficiently high, then an earlier fusion can potentially be used. However if substantial noise is present, a later fusion may be needed in order to "clean the data" within the individual modalities prior to fusing.

The ideal choices are related to the dependencies and degree of complementarity across the modalities as reviewed in section 3.2.3. When strong dependencies exist among the modalities, an early fusion is ideal, as the simultaneous modeling of all modalities will then allow the dependencies to interact and hence strengthen the signal and potentially elicit multivariate signals only existing across modalities. When there are weaker dependencies among the modalities, or when there are latent variables confounding the target signals, an intermediate fusion will have the possibility of identifying the confounder prior to the final decision stage. When the dependencies within each of the modalities are much stronger than the dependencies across modalities, a later fusion should be more effective. The extreme is when the modalities are conditionally independent, where it is obvious that the modalities should be fused at the decision level [Lahat et al., 2015].

Chapter 4

Contributions

4.1 Fusion level

In paper A we investigated integration, or fusion, of multimodal data, with a specific focus on medical data. The multimodal data used in the paper consisted of structural and functional MRI (sMRI and fMRI respectively) scans from the same population of subjects.

The main question of the paper was how to achieve a more optimal fusion level when exactly two modalities comprise the multimodal data set. We claim that the optimal fusion level is related to the dependencies across modalities, hence taking the "ideal" approach to fusion as described in section 3.3.4. We investigated this at three fusion levels, namely early, intermediate and late fusion.

We defined the three fusion levels by the formulations in table 4.1. The complete data set of J modalities is in table 4.1 denoted $u = [v_1, \ldots, v_J]$, where v_j is the data from modality j. h denotes latent variables that could elicit dependencies across modalities, e.g. endophenotype, gender or age.

Table 4.1: Bayesian formulations of early, intermediate and late fusion, as described in paper A.

Early	Intermediate	Late
$p(y\|u) = \frac{p(u\|y)p(y)}{p(u)}$	$p(y\|u,h) = \frac{p(y,h)\prod_{j=1}^{J} p(v_j\|y,h)}{p(u,h)}$	$p(y\|u) = \frac{p(y)\prod_{j=1}^{J} p(v_j\|y)}{p(u)}$ $\Rightarrow$ $\frac{p(y\|u)}{p(y)} = \prod_{j=1}^{J} \frac{p(y\|v_j)}{p(y)} \prod_{j=1}^{J} \frac{p(v_j)}{p(u)}$

It is seen that for the late integration, the posteriors of the full data set $p(y|u)$ and the posterior from the individual modalities $p(y|v_i)$ are in a Bayesian surprise constellation [Itti and Baldi, 2009], where they are placed relative to the prior on y, $p(y)$.

The dependency structures in the data were investigated using a permutation scheme to create and break dependencies across modalities (see figure 4.1). By shuffling the variables among the modalities and splitting them up into new modalities, interdependencies were created across modalities, hence mimicking an early integration data set. Similarly a late integration data set was created

by shuffling the subjects within each group of each modality, thus making the modalities conditionally independent, i.e. potential confounding dependencies were broken across modalities.

The permuted data sets (B and C in figure 4.1) and the original data set (A in figure 4.1) were evaluated on a deep pipeline in early, intermediate, and late configurations (see figure 4.2). In the early configuration, modalities were fused by concatenation as a first step and the fused data matrix was used as input for an unsupervised restricted Boltzmann machine (RBM) [Smolensky, 1986, Hinton, 2002]. The output nodes of the RBM were used as input for the decision layer, which was a logistic regression model. The intermediate configuration kept the modalities separate until after the RBM, where the output from each of these were fused and inputted to the logistic regression. Finally in the late configuration, the modalities were analyzed with completely separate pipelines comprising separate RBMs and decision layers, and only at the end were the probabilistic outputs of a final sigmoid layer compared across modalities to establish a final classification decision.

The results of the baseline data were compared to the extreme cases created using the permutation schemes in order to conjecture on the true dependencies within the modalities.

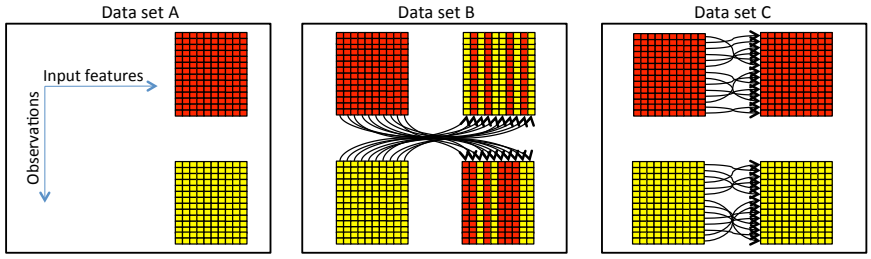

Figure 4.1: Figure from paper A illustration the permutation scheme. Data set A is the original unpermuted data set. In data set B, the features of the two modalities were permuted to increase dependencies across modalities. In data set C, the observations were permuted within the group to create class conditional independence among the modalities.

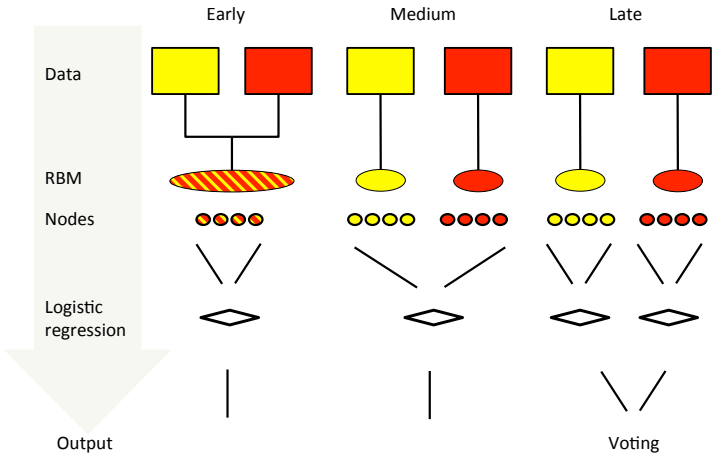

Figure 4.2: Figure from paper A illustrating the analysis pipeline used. Each color denotes a separate data modality. "Nodes" denotes the output of the RBM. The output is either a thresholding of the probabilistic output of the logistic regression, or a combination of the probabilistic outputs of two logistic regression models.

4.1.1 Rewriting the late fusion formula for decision making

The Bayesian formulation of the late fusion not only describes the assumptions of the decision level fusion, it can also be rewritten to be applied as a means of fusing outputs from multiple classifiers trained on individual modalities. This method has previously been described as "the product rule" of combining decisions [Kittler et al., 1998]. The following section is not part of the original publication, but is a rewriting of the late fusion formula to allow for a numerically stable implementation of the decision level fusion method.

Continuing from the Bayesian surprise construction shown in table 4.1, as $\prod_{j=1}^{J} \frac{p(v_j)}{p(u)}$ is independent of y, it is regarded as a normalization. Hence

$$p(y|u) = \frac{p(y) \prod_{j=1}^{J} \frac{p(y|v_j)}{p(y)}}{\sum_{c=1}^{C} p(y_c) \prod_{j=1}^{J} \frac{p(y_c|v_j)}{p(y_c)}}$$

where C is the number of classes.

The function $q(y')$ is introduced as the logarithm of the product of the likelihood and the prior

$$\begin{aligned} q(y') &= \log\left(p(y')\prod_{j=1}^{J}\frac{p(y'|v_j)}{p(y')}\right) \\ &= \sum_{j=1}^{J}\Big[\log(p(y'|v_j))\Big] - (J-1)\log(p(y')). \end{aligned}$$

The maximal value of q for a given input y' is defined as

$$q(y')_{max} = \max_{y'}\left\{\sum_{j=1}^{J}\Big[\log(p(y'|v_j))\Big] - (J-1)\log(p(y'))\right\},$$

and is subtracted from the numerator and denominator, to ensure numerical stability

$$p(y|u) = \frac{\exp\Big(q(y) - q(y)_{max}\Big)}{\sum_{c=1}^{C}\exp\Big(q(y_c) - q(y)_{max}\Big)}.$$

4.2 Transferability of learning efficiency

In paper B we explored the variation in efficiency of learning from data among different models. Models within machine learning differ in how efficiently they learn from data which is expressed by the varying amount of training data needed to reach an asymptotic error of a problem. In that sense, some models are quicker learners compared to other models, also within the same level of complexity. The generative naïve Bayes model has been found to be a quicker learner compared to the discriminative logistic regression model [Ng and Jordan, 2002, Xue and Titterington, 2008]. This relationship was in the present study explored using a learning curve approach to monitor the performance of the individual models as more training data was made available to the model.

We further investigated the possibility of transferring the quick learning properties from the generative model to the discriminative model by augmenting the training set of the discriminative model with synthetic samples generated from the generative model. The performance of the augmented discriminative model was evaluated based on the amount of augmented data and was compared with the two base models using the learning curve framework.

The quick and slow learning properties and the augmentation scheme were initially evaluated on synthetic data, where the properties of the models were investigated with a perspective on the distribution and the covariance structures of the data across classes. Four distributions were evaluated: normal distribution, students t-distribution with df=3, lognormal distribution, and a mixed normal distribution where each class had two latent clusters sampled from four normal distributions with different means and covariance matrices.

From the simulated data we found that when the naïve feature independence assumption is fulfilled, and the data is distributed according to or similar to a normal distribution, i.e. normal, student, and mixed normals, the naïve Bayes model is the quicker learner. When the distribution is lognormal or the covariance matrix is full, the logistic regression model is the quicker learner. Furthermore, it seems that, when the naïve Bayes model is the quicker learner, it is possible to generate synthetic data from the model and augment a logistic regression model with an improved performance as a result.

To validate this finding, the hypotheses were further evaluated in eight real world data sets from the UCI repository [Dheeru and Karra Taniskidou, 2017]. It was found that in seven out of eight data sets, the naïve Bayes model was indeed the quicker learner and in these cases, the performance of the logistic regression model augmented with the synthetic samples, was indeed improved relative to the base logistic regression model. In five out of eight data sets, the

augmented model achieved a significantly better performance compared to both baseline models when only a small amount of data was available. In two data sets, the augmented model performed on par with the baseline models.

When only a small amount of data was available to the models (see figure 4.3), the sample efficiency was computed for the comparison of the best of the two base models (denoted in the paper as the "oracle" model) and the augmented model according to equation 4.1.

$$\text{Sample efficiency} = \frac{N_{tr}^{OR} - N_{tr}^{AM}}{N_{tr}^{AM}}, \tag{4.1}$$

where N_{tr}^{OR} is the number of true training samples needed by the oracle model and equivalently for N_{tr}^{AM} for the augmented model.

A positive sample efficiency means that the oracle model requires more training samples compared to the augmented model, and a negative that the oracle model needs less. For six out of the seven data sets, where naïve Bayes was found to be the quicker learner, the sample efficiencies were positive for all experiments where the training set of the augmented model had been augmented with 150 synthetic data samples or more. Sample efficiencies well above 200% were seen.

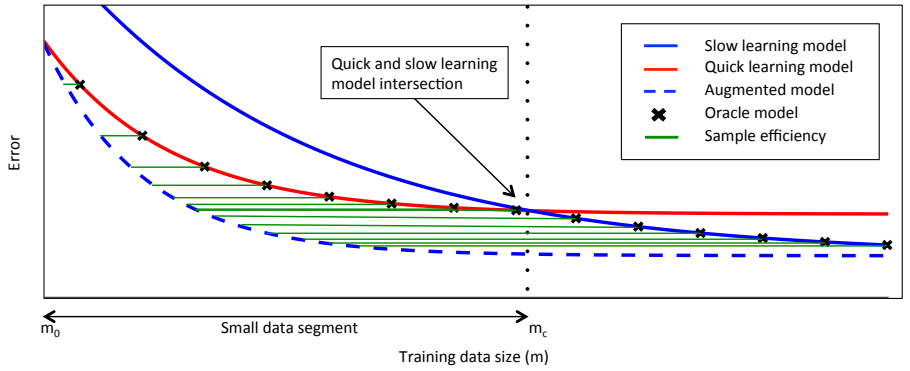

Figure 4.3: Concept figure from paper B illustrating the learning curve approach, including the small data segment and sample efficiency. The augmented model is a variant of the slow learning model where the training data is augmented with synthetic data generated from the quick learning model. The oracle model is the better of the slow and the quick learning models at the given training data size.

4.3 Diagnostic classification accuracy

In paper C, the possibility of accurately classifying schizophrenia patients from healthy controls based on multimodal data acquired from patients with ANFE schizophrenia was explored. Additionally, the possibility of predicting symptom remission in patients at a six week follow-up examination was evaluated.

The multimodal data investigated in this study was comprised of a range of cognitive tests, electrophysiological tests, sMRI data and DTI data. Forty-six patients and fifty-six healthy controls were included from cohort C acquired at CINS . Subjects where all variables from a modality (e.g. all the cognitive tests) were missing were excluded. At the six week follow-up, twelve patients had either dropped out or were excluded due to missing data or other reasons, leaving thirty-four patients for the remission prediction study. Of the thirty-four patients, eleven were in remission after six weeks according to the Andreassen criteria [Andreasen et al., 2005].

The analysis pipelines for both studies were build using nested cross validation (CV). The outer loop was constructed to obtain a generalization error and consisted of 100 stratified random-subsampling splits with 1/3 of the data in each split for testing and 2/3 for training. Within the loop, training data was standardized to zero mean and unit variance, missing variables were imputed using nearest neighbor. Following preprocessing, nine configurations of six algorithms were trained on the training data. For models with hyperparameters that needed optimizing, this was done in an inner five-fold cross validation loop. For models without hyperparameters or where hyperparameters were found heuristically, a backwards elimination feature selection was implemented based on a model of the same type and using the inner CV loop.

The analyses were conducted first unimodally and each modality that appeared to be able to perform the classification task was fused with the remaining modalities in an early fusion manner. To evaluate the significance of the classification performance, permutation testing was performed using 1000 random permutations of the true labels.

For the diagnosis prediction study, the nine configurations of algorithms when trained on the cognitive data yielded accuracies ranging from 60% to 69%. The baseline accuracy was 56%. The accuracies were all found to be better than chance at a statistically significant level ($p < 0.05$). Of the four modalities investigated, only cognition had any unimodal prediction power. When cognition was fused with the remaining individual modalities and combinations of modalities, prediction performance never improved above the unimodal prediction accuracy.

For the prediction of symptom remission, none of the classifiers could predict the outcome based on any of the input modalities at a level greater than chance.

Table 4.2: Training and test accuracies from paper C. The tested methods were naïve Bayes (nb), logistic regression (lr), L1 regularized logistic regression (ls), SVM with a linear kernel (svml), SVM with a radial basis function (RBF) kernel with heuristic identification of parameters (svm), SVM with an RBF kernel with optimized parameters (svmo), decision tree (tree), random forest (rf), and auto-sklearn (ask).

Training, class balance: 0.56									
mdlty	nb	lr	ls	svml	svm	svmo	tree	rf	ask
Cog	0.81	0.90	0.84	0.76	1.00	0.86	0.91	1.00	0.91
EEG	0.71	0.73	0.58	0.65	0.94	0.52	0.92	1.00	0.92
sMRI	0.88	1.00	0.64	0.97	0.99	0.41	0.94	1.00	0.94
DTI	0.75	1.00	0.59	0.95	1.00	0.58	0.93	1.00	0.92
Test, class balance: 0.56									
Cog	0.68	0.61	0.69	0.60	0.64	0.67	0.61	0.69	0.67
EEG	0.53	0.53	0.55	0.55	0.52	0.54	0.55	0.56	0.53
sMRI	0.51	0.50	0.54	0.52	0.53	0.53	0.49	0.51	0.51
DTI	0.53	0.53	0.55	0.54	0.53	0.54	0.52	0.53	0.51

4.3.1 Post-hoc application of data augmentation framework from paper B

After publication of paper C, the framework developed in paper B was evaluated on the cognitive modality. Initially, a learning curve analysis, as developed in paper B was conducted to investigate the learning efficiencies of naïve Bayes and logistic regression on the cognitive modality used paper C (see figure 4.4), however using complete case analysis instead of imputation. This showed that naïve Bayes was the quicker learner, and furthermore that the augmentation scheme seemed to improve the performance of logistic regression beyond both baseline models.

To validate the feasibility of the augmentation approach, the approach was evaluated in the pipeline developed for paper C. The fitting of the naïve Bayes model and subsequent augmentation was implemented following the imputation of missing variables. From the a priori learning curve analysis, augmentation with 200, 400, and 800 samples seemed to perform equally well. As augmentation with 200 synthetic sample yields less computational burden this was chosen

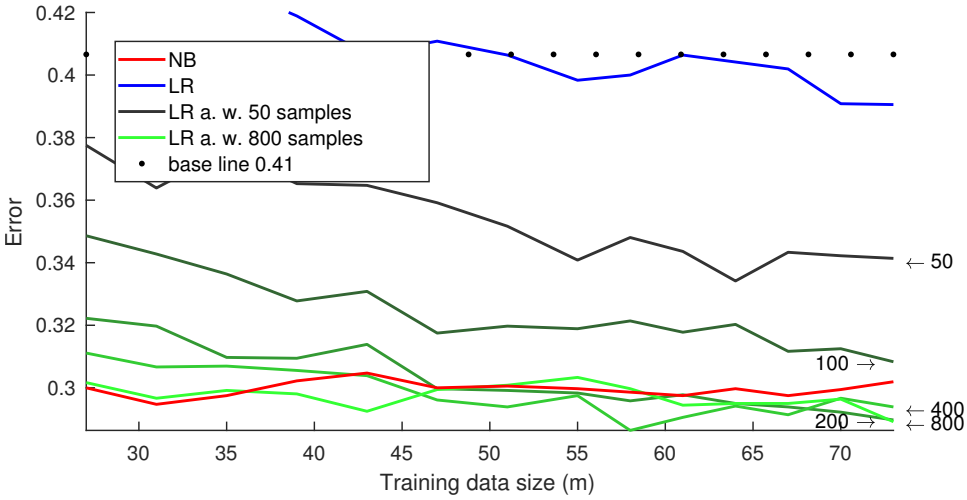

Figure 4.4: Learning curve analysis of the cognitive data, where observations with missing data were excluded (i.e. complete analysis - see section 3.2.2). Naïve Bayes (NB) in red is compared with logistic regression (LR) in blue. Furthermore, logistic regression models where the training set was augmented with 50, 100, 200, 400, 800 samples are shown in green colors from dark to bright (LR augmented with X samples). Baseline is different from what is shown in table 4.2 due to the complete case analysis approach.

for the subsequent analysis. Ideally the choice of the number of generated samples should not have been based on an a priori analysis using the full data set, but rather be determined using cross validation. The pipeline from paper C was employed both with backwards elimination feature selection (FS) and without (No FS).

The results from running the pipeline from paper C are shown in table 4.3, and compared with the performances of the models as reported in paper C (i.e. as in table 4.2 for cognition). It is apparent that most of the models that showed a lower performance in paper C are improved to the level of the better performing models by the augmentation approach. The decision tree is an exception, where only minor a improvement is obtained. These results are discussed in section 5.3.

Table 4.3: Test accuracies from training the classifiers on the cognitive modality augmented with 200 synthetic samples generated via the method in paper B. The augmentation scheme produced as good or slightly better performance compared to the backwards elimination feature selection used for the non-optimized models. Methods tested were naïve Bayes (nb), logistic regression (lr), L1 regularized logistic regression (ls), SVM with a linear kernel (svml), SVM with a radial basis function (RBF) kernel with heuristic identification of parameters (svm), SVM with an RBF kernel with optimized parameters (svmo), decision tree (tree), random forest (rf). FS and No FS is whether the backwards elimination feature selection is included (FS) or not included (No FS). Standard error of the mean, as used for error estimates of CV error in [Hastie et al., 2009], is calculated as $\sigma_{\text{acc}}/\sqrt{N_{\text{CV splits}}}$, where σ_{acc} is the standard deviation of the accuracies from the CV split and $N_{\text{CV splits}} = 100$ in this study.

	Test, class balance: 0.56							
Data	nb	lr	ls	svml	svm	svmo	tree	rf
No FS	0.69	0.69	0.69	0.69	0.67	0.69	0.63	0.69
FS	0.69	0.68	0.69	0.68	0.67	0.68	0.64	0.69
As in paper B	0.68	0.61	0.69	0.60	0.64	0.67	0.61	0.69
Standard error of the mean of the accuracies								
No FS	5.8e-3	6.7e-3	6.2e-3	6.2e-3	6.1e-3	5.9e-3	7.9e-3	6.6e-3
FS	6.2e-3	6.6e-3	6.2e-3	6.5e-3	6.2e-3	6.8e-3	6.8e-3	6.1e-3
As in paper B	6.8e-3	7.4e-3	6.6e-3	6.3e-3	7.1e-3	7.2e-3	8.0e-3	6.5e-3

Chapter 5

Discussion and directions for future studies

Diagnostic support for psychiatric diseases like schizophrenia is a complex and multimodal problem. In the present thesis, different challenges regarding the analysis and fusion of multimodal medical data have been investigated. Furthermore, clinical data from antipsychotic naïve first episode patients has been analyzed. First in a unimodal approach with a focus on diagnostic classification and on prognostic classification. Secondly, the data was analyzed in a multimodal approach, also with a diagnostic classification aim.

Using the papers as starting points the individual studies will initially be discussed separately. The discussions of each topic will then be expanded, as compared to the papers, and finally put into a common context. Especially, the discussion of the paper on accuracy of diagnostic classification algorithms (paper C) will be elaborated. The paper containing the study has a clinical focus, and as such the machine learning part has not been thoroughly discussed.

5.1 Fusion level

In paper A we presented an approach to investigate dependencies among two modalities of a multimodal data set for the purpose of determining the optimal level for data fusion.

The approach is composed of two permutation schemes, each of which is designed to elicit different dependencies in the resulting permuted data set.

The first scheme, yielding data set B in the paper (see figure 4.1), attempts to inflate inter-modality dependencies by shuffling features from the two modalities and splitting them up again into two new pseudo modalities of the same dimensions as the original modalities. This approach builds on the assumption that there are dependencies among the variables within a modality which are potentially stronger than dependencies across modalities. The permutation of features across modalities should be indifferent to the early fusion method. However, intermediate and late fusion should be affected, which was also reflected in the predictive performances of the models. Early fusion achieved a similar performance as in the unpermuted data set, and the performances of the intermediate and late models were worse compared to the unpermuted data set.

If, however, there are only a few informative variables within each modality, while the rest of the variables are uninformative or there is an imbalance between the dimensionalities of the modalities, the potential for establishing increased inter-modality dependencies is limited. Higher variance in the prediction accuracies across the instantiations of the permuted datasets are thus likely to occur. An analysis pipeline with sparsity in the input features would potentially mitigate this issue. As would a feature reduction step prior to permutation (as done in the conceptual study in the paper).

The other permutation scheme (yielding data set C in the paper) permutes the subjects within the groups so as to break inter-modality dependencies. This permutation scheme should hence be indifferent to the late integration (apart from potential effects of batch learning), but could have an effect on the early and intermediate fusion models, given that the data is not conditionally independent. The results in the paper showed that the performance of the late fusion model was indeed the same for the unpermuted and the intra-group permuted data set (data sets A and C in the paper).

The two permutation schemes create the extreme cases, i.e. cases ideal for early and late fusion, respectively, thus defining a fusion level spectrum.

In order to investigate the integration level we implemented a deep pipeline

which would allow us to fuse the data at various levels, i.e. at an early, intermediate and a late decision level. The choices regarding this pipeline of course affected the results. The interpretation of the permutation schemes should hence be seen in the light of the choices made for the pipeline. Furthermore, the data sets, which were applied in the evaluation of the method were already processed using matrix decomposition separately for each modality, meaning that an earlier data fusion was not possible on this data set. Following the terminology presented in figure 3.1, the earliest possible fusion with this data set, is closest to "True fusion using high level features" [Lahat et al., 2015] or "Feature level fusion" [Atrey et al., 2010]. The framework has hence not been evaluated on the true extremes of the data fusion spectrum as defined in the literature (see figure 3.1 for clarification on this).

In an effort to construct the permutation scheme such that it would be easy to apply, we chose to confine the discussion in the paper, to discrete levels of fusion - i.e. "early", "intermediate" or "late" fusion. However, it is likely that the fusion level scale would be better modeled as a continuum spanning from the very early raw data fusion to a late decision level fusion. This interpretation will be included in the present discussion, with the framework established by the two extremes of the spectrum relative to the data set available.

The results from the analyses showed that the late decision level fusion provided the best performance compared to the early and intermediate fusion steps. This indicates that the data sets from the two modalities, essentially are conditionally independent. This conclusion is however also conditioned on the ability of the analysis pipeline to identify the signal from the much higher dimensional space which is a consequence of early fusion. The worse performance of the late fusion model when trained on data set B compared to the unpermuted data set further indicates that the intra-modality dependencies are stronger than the inter-modality dependencies. Further support for the hypothesis of conditional independence is found in the results of the early (and intermediate) fusion of the intra-group permuted data set (data set C in the paper). They show that, despite of the broken link between the two modalities - i.e. the fMRI data of one subject is fused with the sMRI data of another subject within the same group, it is still possible to obtain an accuracy comparable with the unpermuted data set. There even seems to be an improvement in the performance of the intermediate fusion model, which could indicate that, with some existing noise covariances broken, it is easier for the intermediate level model to identify remaining latent confounders prior to fusing.

Further analyses that could be done in this framework include an expansion to allow analysis of more modalities than two. This will, however, be limited by an issue regarding the early fusion permutation scheme (data set B), where, as the number of modalities increases, fewer variables from each modality will

be combined in each new pseudo modality, hence diluting the inter-modality strength and also the intra-modality dependencies. There is hence a limit for the scheme which is bounded by the number of modalities and the dimensionality of each modality - and more explicitly the number of informative features in each modality.

For the intra-group permutation scheme, there is a similar limit which is related to the number of modalities and the number of subjects in the smallest group N_{min}. By permuting subjects, such that the overall pattern of subject combinations is unique for a given permutation, the number of possible permutations is $N_{min}!$, which will not constitute an issue for sample sizes of general usability in machine learning.

As shown previously, rewriting of the Bayesian theorem to accommodate the conditional independence hypothesis of multiple modalities further acts as an actual means of fusing modalities in a decision level fusion (see section 4.1.1). Given that the assumptions regarding the data holds, and that the posterior probabilities computed for each modality $p(y|v_j)$, are proper calibrated posterior probabilities representative of each modality as input, this should be an optimal way to perform late fusion. There are hence two main uncertainties in the viability of the fusion method.

Firstly, the modalities might not be conditionally independent. This would entail that the fusion is naïve towards dependencies among modalities, hence potentially suboptimal. Further studies should be conducted to investigate the effect of the inaccuracy of this assumption. Relying on the studies of naïve Bayes, which implements similar assumptions of feature conditional independence, as opposed to modality conditional independence, it is however, likely that the integration method would work even if the conditional independence assumption is void [Zhang, 2004].

The second uncertainty relates to the posterior probabilities which are obtained from training classifiers on the individual modalities. This hence requires that the individual classifiers are in fact properly calibrated in their posterior probabilities. Kittler and colleagues found that the fusion method, which they call "the product rule", is extra sensitive to estimation errors, which is the error a model makes in predicting the true posterior probability [Kittler et al., 1998], hence confirming the need to ensure that only well-calibrated probabilistic classifiers are used for decision level fusing using this method.

In cases where more than two classes are present in a data set and where the classes might not be mutually exclusive. Dempster-Shafer (DS) theory allows assigning probability to groups of classes as well as individual classes. A scenario where this might be relevant is in a general psychiatry data set with data

from patients with a range of psychiatric diagnoses. There is evidence that the symptoms of different diagnoses are in fact overlapping, and that they are potentially influenced by the subjectivity of the psychiatrist giving the diagnosis (as already reviewed in section 2). In this scenario, DS theory could allow subjects to belong to multiple groups of diagnoses, hence potentially illustrating an inter-disease subtype.

5.2 Transferability of learning efficiency

In paper B we focused on the issue of learning from small data samples by exploring the difference in efficiency with which models learn from data. Specifically, we explored the learning efficiencies of the generative-discriminative model pair - naïve Bayes and logistic regression, where naïve Bayes has been found to be the quicker learner in most cases [Ng and Jordan, 2002]. We verified the learning efficiencies of the two models through empirical studies on simulated and real data. We further explored the potential to transfer the quicker learning abilities to the slower learning logistic regression by creating synthetic data from the generative naïve Bayes. We found that when naïve Bayes is indeed the quicker learner, it is possible to decrease the amount of true data needed for logistic regression to achieve asymptotic performance by augmenting the training data with synthetic data generated from the naïve Bayes model.

From the simulated data it was found that naïve Bayes is indeed the quicker learner, when the assumptions of the model are fulfilled or close to being fulfilled. That is, when data is distributed with a normal, a student-t or mixed normal distribution with a common or individual diagonal covariance matrices. In the other cases, logistic regression was much faster. It has however been found that naïve Bayes can provide an accurate prediction even when these assumptions are void [Domingos and Pazzani, 1997]. Domingos and Pazzani further find that the degree of dependencies among variables in a data set is not predictive of the degree of errors made by the naïve Bayes model, hence concluding that these dependencies are not necessarily essential to uncover in order to improve the performance of the classifier. Zhang identified a special case where, despite the naïve assumptions being void, naïve Bayes can still be an optimal classifier. These are cases where existing dependencies among variables cancel each other out. This could give pointers to where it might be beneficial to augment training data with synthetic data from naïve Bayes, however further studies should be conducted to investigate this.

In our results we found that for one data set, naïve Bayes was not more data efficient compared to logistic regression, namely for the liver disorders data set.

Therefore, the synthetic data generated from the naïve Bayes model did not improve the learning efficiency of the logistic regression model. Considering the fact that naïve Bayes was not the more efficient model in relation to the results from the simulated data, could suggest that the liver disorders data is not distributed according to the assumptions of conditional independence made by naïve Bayes. Domingos and Pazzani also investigated the liver disorders data set and found that better predictive performance was achieved by categorizing the variables and training a multinomial naïve Bayes model instead of a Gaussian based model as applied in paper B. This suggests that it is the assumption of normal distributed variables that is void in this data set, and not necessarily the independence assumption.

In paper B it was found that for most cases, the predictive performance of logistic regression could be improved by augmenting the training data with synthetic data generated from the naïve Bayes model. It is however not clear what the optimal number of generated samples are for a given data set and why, though it is clear that it varied among the evaluated data sets. For some data sets it was related to how accurate the base models (naïve Bayes and logistic regression) were, and how large the gap was between their respective predictive performances, i.e. by augmenting the training set of logistic regression with synthetic data, the predictive performance was interpolated between the baseline models according to how much synthetic data, the training data was augmented with. In a few other cases the trend was however that more synthetic data was always better. In these cases the augmented data seemed to boost the performance of logistic regression beyond the performances of both base models. This relationship should be investigated in a future study.

An extension to the question of what the optimal number of generated samples is, is why the learning efficiency of the models is different from the synthetic data compared to the true data. The synthetic data, seemed in most cases to be providing a lower learning efficiency. Though the time for training a model is increased with larger data sizes, it is still only a fraction of the cost compared to obtaining more true data. It would however be interesting to investigate the reason for this quality difference, and potentially identify how higher quality synthetic data could be generated. A potential pragmatic approach, which however is less efficient, would be to implement a rejection option to the data as done in general adversarial networks (GANs). Another approach could be to assign weights to the true data and the synthetic data, and potentially vary the weights, e.g. in a boosting manner [Hastie et al., 2009].

The quicker learning properties has been found for the multinomial naïve Bayes as well [Ng and Jordan, 2002, Xue and Titterington, 2008]. The possibility of using the multinomial naïve Bayes in the proposed data augmentation framework should be investigated in future studies.

5.3 Diagnostic classification accuracy

In paper C, we investigated the potential of distinguishing a group of ANFE schizophrenia patients (46) from a group of healthy controls (58) based on data from cognitive tests, electrophysiological measures, structural MRI and DTI data.

To investigate the diagnostic classification potential for this data, we selected a range of machine learning models, which we trained on the data to predict the diagnosis. The models were chosen with inspiration from other studies where machine learning methods have been used to analyze neuropsychiatric data, but also with regard to models which have previously been found to be able to learn from relatively small amounts of data (e.g. naïve Bayes).

Models which had no inherent parameters to optimize or where parameters were chosen heuristically or using default values, a backwards elimination feature selection was implemented using the specific model as the method for evaluating feature viability. This was done to let the models have a more equal possibility to learn from the relatively high dimensional data.

The results from training the machine learning models to predict the diagnosis showed that all models had significant predictive performance when based on the cognitive modality, however when trained on data from the electrophysiological modality or any of the neuroanatomical modalities, no predictive capabilities were found. When training the models to predict symptom remission within the patient group, none of the models showed any predictive capabilities.

A few of the models, e.g. logistic regression, were found to overfit to the training data, particularly for the high dimensional neuroanatomical modalities, seen by training accuracies close to 100% (result not shown in paper, but shown in table 4.2). Especially logistic regression included all or close to all variables after feature selection for the higher dimensional data sets. Comparatively it was seen that the L1 regularized logistic regression, which performed better for the cognitive modality, exhibited a lower training accuracy (see table 4.2), and also seemed to identify the same variables as the univariate test (see figure 3 in paper C), hence suggesting that it did not overfit. There was, however, only a relatively small improvement of the generalization error gained from the models that did not overfit, indicating that the chosen approach for optimizing the models was valid for the give data sets.

Looking more into the results of the feature selection, it was found that the features that showed lower p-value in the univariate tests, had a tendency to be included more often in the feature selection scheme (see figure 3 in paper

C). This could indicate that the strongest signals in the data are the univariate signals, and that there is not enough power in the multivariate relations in the data to model these efficiently with the given sample size.

The generalization errors obtained from training the models on the non-informative modalities of the diagnostic test and on all the modalities in the remission prediction, showed that the mean accuracy of several models was below the chance accuracy (0.56), showing mean accuracies closer to 0.5. This scenario can occur if two clusters have the same mean values and the classifier attempts to split the groups with a linear classification boundary running through this mean value. The accuracy will then become 0.5 instead of the balance value between the two potentially skewed distributed classes. Whether this was what actually happened in these analysis could be a topic for further investigation.

As only the cognitive modality was found to have predictive capabilities, the multimodal studies were conducted from the hypothesis that the noise patterns in the different modalities could potentially be complimentary, hence potentially improving the signal found in the cognitive modality. The results showed a small decline in the predictive performance of the multimodal models compared to the unimodal analysis and are hence not in favor of this hypothesis. It is however likely that the task was made harder by adding non-informative variables to the informative cognitive modality. The fact that the models still seemed to predict the diagnoses above chance support the view that the amount of data was sufficient for the models to learn from the data.

The modalities were fused at a relatively early level - that is using what Lahat and colleagues denote true fusion using multivariate features [Lahat et al., 2015] (see also figure 3.1). The fusion method was a simple concatenation of the data sets from each modality. This of course leaves the question of whether it might have been beneficial for the analyses to, e.g use a data driven matrix decomposition method, to potentially elicit more condensed features. This could be done either on the individual modalities or used as the fusing mechanism.

The decomposition of individual modalities using e.g. PCA or ICA could have replaced the feature selection step in the processing pipeline. This was a design choice, where feature selection was favored, as this allowed for a better comparison between models with inherent feature optimization and models without.

Decomposition across modalities as a means of fusion, where the decomposition method could be used to identify common latent variables, would be an interesting topic for a future analysis. However, the relatively small sample size is likely to limit the possibility of the models to fit the data without strong regularizing assumptions.

The unimodal results did not support a later fusion, since only the cognitive modality was actually informative. Thus making a decision level fusion meaningless.

Despite the negative results from electrophysiology, sMRI and DTI, considering the heterogeneity of schizophrenia, it can not be concluded that these modalities do not have the potential to assist in diagnosing patients suffering from a subtype of schizophrenia. Given the potential bias of the analyzed sample and the relatively small data size, a general dismissal of the analyzed modalities for the use in ANFE schizophrenia should not be the conclusion based on the negative results from this study alone. It should, however, encourage further studies in independent samples to verify these findings.

After the publication of paper C, the augmentation approach from paper B was evaluated on the cognitive data alone, as this was the only modality with any predictive capabilities. The models trained on the augmented data sets were found to improve performance of many of the evaluated models - even if the very time consuming feature selection step was omitted. The number of generated samples was set to 200 based on the results of an initial learning curve analysis on the cognitive data using the framework of paper B. This way of identifying the optimal number of generated samples is clearly a violation of keeping the test set "locked away". This analysis should have been done in an inner cross validation loop setting, where the feasibility of the approach could initially be confirmed for the given data set. More specifically, in an inner loop it should 1) be determined if naïve Bayes indeed performs better than logistic regression, and 2) the optimal number of generated samples should be identified via a grid search.

The decision tree showed a smaller improvement from the augmentation scheme compared to the other models, and contrary to many of the other models, it showed a small decrease in performance when feature selection was omitted. This could indicate that different models require different numbers of synthetic samples. The model sensitivity to the number of generated samples should be evaluated in a future study.

The fact that the performance of all the models seemed to be improved by the augmentation approach, indicates that the approach is indeed model agnostic, and can hence has the potential to function as a general data augmentation method for all data types.

5.4 Perspectives

In the present thesis, multimodal medical data, particularly data from antipsychotic naïve first episode schizophrenia patients has been investigated. Two approaches of maximizing retrieval of potential information using machine learning methods from multimodal medical data have been investigated. We proposed frameworks to investigate how to best fuse multimodal data and how to potentially fit more complex and slower learning models to small data sets. Furthermore, a study on diagnostic classification accuracy has been conducted using data from ANFE schizophrenia patients, where we found that contrary to our own hypotheses we could not identify the patients from the controls based on neither electrophysiological nor neuroanatomical modalities, but only on cognitive measurements.

Theoretically, regardless of the actual dependencies across modalities, an ideal data driven early fusion method would be able to identify the relevant diversities of the modalities, such that an optimal fusion is achieved. Poor SNR, however often prohibits this in real world data, which leads to pragmatic fusion level choices. On the other hand, it may be beneficial to actually be aware of the dependencies in the data at hand, such that unnecessary constraints on the models can be avoided, and instead an informed choice can be made about the fusion level. In paper A we developed a method to help making this informed choice.

In paper B we investigated how to potentially enhance the learning efficiency of a discriminative model, via data augmentation. The approach is agnostic to the format of the data, meaning that both image data and non-image data can be augmented using the approach. In all data sets where the generative model was an efficient learner, it was possible to increase the learning efficiency of the discriminative model.

In an attempt to provide guidelines for future approaches to the conundrum of fusing multimodal medical data, it is evident that the obstructions in the data we analyzed were greater than anticipated. Despite univariate group differences having been identified for several measures from individual modalities, the effect sizes were not large enough and neither were the multivariate and multimodal diversity, leaving the current state of single subject diagnostic classification using machine learning methods for the given data set at an impasse.

The proposed framework for investigating integration level of multimodal data is dependent on there being relevant signals in all modalities staged for fusion. As this was not the case in paper C the framework could not be used for this.

The results from applying the augmentation scheme to the cognitive modality encourages its specific use in this data set. Additionally the developed data augmentation scheme seems in general applicable to machine learning analyses of small data sets, particularly of non-image data, where the selection of data augmentation methods are sparse.

Appendix A

Paper A

Testing Multimodal Integration Hypotheses with Application to Schizophrenia Data

---

Included in the proceedings of the 5th InternationalWorkshop on Pattern Recognition in Neuroimaging (PRNI), 2015.

[Axelsen, M. C., Bak, N., and Hansen, L. K. (2015). Testing Multimodal Integration Hypotheses with Application to Schizophrenia Data. *In 2015 International Workshop on Pattern Recognition in NeuroImaging (PRNI)* (pp. 37-40). ieeexplore.ieee.org.]

Testing Multimodal Integration Hypotheses with Application to Schizophrenia Data

Martin C. Axelsen*†, Nikolaj Bak†‡ and Lars K. Hansen*
*Cog-Sys - DTU Compute, Technical University of Denmark
Kgs. Lyngby, Denmark, Email: maxe@dtu.dk
†Center for Clinical Intervention and Neuropsychiatric Schizophrenia Research (CINS),
Psychiatric Centre Glostrup, Mental Health Services, Capital Region, Denmark
‡Center for Neuropsychiatric Schizophrenia Research (CNSR),
Mental Health Services, Capital Region, Denmark

***Abstract*—Multimodal data sets are getting more and more common. Integrating these data sets, the information from each modality can be combined to improve performance in classification problems. Fusion/integration of modalities can be done at several levels. The most appropriate fusion level is related to the conditional dependency between modalities. A varying degree of inter-modality dependency can be present across the modalities. A method for assessing the conditional dependency structure of the modalities and their relationship to intra-modality dependencies in each modality is therefore needed. The aim of the present paper is to propose a method for assessing these inter-modality dependencies. The approach is based on two permutations of an analyzed data set, each exploring different dependencies between and within modalities. The method was tested on the Kaggle MLSP 2014 Schizophrenia Classification Challenge data set which is composed of features from functional magnetic resonance imaging (MRI) and structural MRI. The results support the use of a permutation strategy for testing conditional dependencies between modalities in a multimodal classification problem.**

I. Introduction

Schizophrenia is a complex disorder with a very heterogenous symptomatology [1]. To assist diagnosis many quantitative techniques including neuroimaging have been proposed although no modality has solved the diagnosis problem yet. Hence, new proposed diagnostic tools typically face a complex multimodal decision challenge. Combining data from several modalities in a classification pipeline is not trivial as this can be done at several levels. Multimodal decision problems are in fact relevant to several fields, and three different levels of integration are applied ranging from the early integration of modalities in data level fusion towards an intermediate integration at the feature or representation level, and finally a late integration often named decision level [2][3]. Hybrid integration is also discussed where data is fused at different levels [2][4].

In a review by Sui et al [5], several multivariate methods of early to intermediate fusion of brain imaging data are discussed. The earlier fusion levels (denoted data fusion) are chosen for the review[5] as these allow for access to potential joint information between the several modalities where later fusion (denoted data integration) preclude the decision model to explore such information.

Later fusion may however lead to simpler models with less parameters to be inferred, hence, potentially less data overfitting and reduced computational complexity. An additional benefit from a late fusion scheme, is that it will be possible to obtain results in cases where one or more modalities are missing. This is of particular interest in complex diagnostic problems where patient conditions can preclude acquisition of data [6].

Conditional independencies can be explored for the identification of the right level of integration for a given data set. Let $\mathbf{y}$ be the relevant decision label, $\mathbf{h}$ latent variables, and $\mathbf{u} = (\mathbf{v}_1, \mathbf{v}_1, ..., \mathbf{v}_J)$ be observed multimodal data for J modalities. Decision theory tells us that the error rate (or more generally the expected loss) is minimized when decisions are based on the posterior probability $p(\mathbf{y}|\mathbf{u})$. To guide our inference procedures, we use Bayes theorem to rewrite

$$p(\mathbf{y}|\mathbf{u}) = \frac{p(\mathbf{u}|\mathbf{y})p(\mathbf{y})}{p(\mathbf{u})}. \tag{1}$$

If the only information shared between the modalities is the decision label, i.e., the modalities are conditionally independent, we obtain

$$p(\mathbf{y}|\mathbf{u}) = \frac{p(\mathbf{y})\prod_{j=1}^{J} p(\mathbf{v}_j|\mathbf{y})}{p(\mathbf{u})}. \tag{2}$$

This allows for a multiplicative combination scheme for the 'Bayesian surprise' as defined in [7], hence, by normalization, of the posterior probability of interest

$$\frac{p(\mathbf{y}|\mathbf{u})}{p(\mathbf{y})} = \prod_{j=1}^{J}\left[\frac{p(\mathbf{y}|\mathbf{v}_j)}{p(\mathbf{y})}\right]\left[\frac{\prod_{j=1}^{J} p(\mathbf{v}_j)}{p(\mathbf{u})}\right]. \tag{3}$$

This is a late fusion scheme as we essentially combine J independent inference pipelines to reach the optimal decision function $p(\mathbf{y}|\mathbf{u})$.
Now, such conditional independence with respect to the diagnosis, $\mathbf{y}$, may be a too strong assumption. Other relevant features known or unknown could create dependencies between the modalities, such as age, gender, or endo-phenotypes. If we denote these variables by $\mathbf{h}$, then by a similar argument, modality independence conditioned on labels and

latent variables, i.e., $p(\mathbf{u}|\mathbf{y},\mathbf{h}) = \prod_{j=1}^{J} p(\mathbf{v}_j|\mathbf{y},\mathbf{h})$, leads to an 'intermediate' level fusion scheme, where a first set of independent pipeline modules identify $\mathbf{h}$.

As the salient dependency structures are unknown, we may simply as a first step (A) explore the relative merits of early, intermediate, and late fusion architectures. As a step towards more detailed understanding, however, we here suggest two additional permutation steps to test dependencies in a given data set. We propose a permutation step designed to increase inter-modality dependency and a step to remove them. The permutation schemes are illustrated in Fig 1. In proposed step (B) we permute the measurement variables to create a randomized set of "pseudo-modalities" each having the same number of variables as in the original measurements. By this operation, possible within modality dependencies are transformed to become dependencies between modalities. If we adapt an early fusion model on such permuted data we clearly expect no difference compared with the early fusion model for case (A). However, when training intermediate or late fusion models a performance drop will inform us that the biases introduced in these simpler architectures are too strong for the task. Our final permutation strategy (C) tests the assumption of conditional independence on labels $\mathbf{y}$. We create a new data set in which we permute the sample indices on the individual modalities among data within a specific $\mathbf{y}$ group, i.e., if we consider a binary decision problem, we randomly mix modality subsamples within the two label groups. Thereby we create a new sample in which any other dependency than that induced by $\mathbf{y}$ has been removed. Under the late fusion hypothesis this step should not decrease performance relative to late fusion in case (A).

II. Materials and methods

A. Data

Data was obtained from the Kaggle website (https://www.kaggle.com): "The MLSP 2014 Schizophrenia Classification Challenge" (partially describe in [8]). It consists of 378 features from a functional magnetic resonance imaging (fMRI) paradigm and 32 features from a structural MRI (sMRI) scan from 86 observations (40 schizophrenia patients and 46 healthy controls). The Kaggle challenge was to classify patients vs. controls (binary classification). Only the labeled part of the dataset was used.

B. Pipeline

In order to investigate the dependencies between modalities, three levels of fusion (early, intermediate, late) were tested. First, input feature selection was performed with the filter method [9]. The ten lowest ranking input features (judged by p-value) from each modality were included. The number of included features were a compromise between the dimensions of the original modalities and our aim to treat the modalities at approximately same footing.

The main non-linear processing step, designed to infer relevant latent features $\mathbf{h}$, consisted of a restricted Boltzmann machine (RBM). The decision step was performed with logistic regression on the binary nodes and the group, see Fig. 2. The restricted Boltzmann machines were modified from the implementation in [10] to accommodate Gaussian distributed visible units. Contrastive divergence as introduced by Hinton [11] is used for learning.

Assuming that the variance of the input data is 1, the updates for the visual and hidden units are then given by

$$\mathbf{h}_{data} = \sigma(\mathbf{b} + \mathbf{v}_{data}\mathbf{w}^T) > 0.5 \tag{4}$$

$$\mathbf{v}_{recon} = \mathbf{a} + \mathbf{h}_{data}\mathbf{w} + \epsilon \tag{5}$$

$$\mathbf{h}_{recon} = \sigma(\mathbf{b} + \mathbf{v}_{recon}\mathbf{w}^T) > 0.5 \tag{6}$$

where ϵ is unit variance Gaussian white noise, σ is a sigmoid function, $\mathbf{a}$ is the bias for the visual units, and $\mathbf{b}$ is the hidden unit bias. Thus, the updates for each variable are

$$\widehat{\mathbf{w}} = \beta\widehat{\mathbf{w}}_{-1} + \alpha((\mathbf{vh})_{data} - (\mathbf{vh})_{recon}) \tag{7}$$

$$\widehat{\mathbf{a}} = \beta\widehat{\mathbf{a}}_{-1} + \alpha(\mathbf{v}_{data} - \mathbf{v}_{recon}) \tag{8}$$

$$\widehat{\mathbf{b}} = \beta\widehat{\mathbf{b}}_{-1} + \alpha(\mathbf{h}_{data} - \mathbf{h}_{recon}) \tag{9}$$

where α is the learning rate, β is the momentum and the subscript (-1) denotes the update of the variable from the previous iteration.

Weights were initialised randomly, the bias for the visual units was initialised as $\log\left(\frac{40/86}{1-40/86}\right)$, and the bias for the hidden units was initialised as zeros, all as recommended in [12]. The values for the learning rate ($\alpha = 10^{-2}$), momentum ($\beta = 2^{-1}$) and batch size (10 samples) were found according to this guide as well. Logistic regression was finally used for computing posterior probability outputs with the RBM nodes as input and the group as label.

Grid searches were done to assess the optimal number of hidden nodes in each integration procedure. This was done to ensure that possible differences in performance could not be attributed to model complexity or number of parameters alone.

The aim of the present study was to explore more formally the effects of possible conditional independence between modalities, given the group. Therefore, based on the original data set (data set A), two additional data sets were created as described above by permutations of input features (B) and among observations (C), respectively (see Fig. 1). In the first permutation strategy (data set B), the features from the two modalities, fMRI and sMRI, were mixed, so two new pseudo modalities were created based on the original data. The grid search for this combination was restricted to equal number of hidden nodes by symmetry. In the second permutation strategy (data set C) the sample indices within a given label group were randomly permuted for each modality.

C. Crossvalidation

The performances of the three levels of integration in each data set were estimated with an 8 fold cross validation procedure on the entire pipeline after preprocessing (feature selection). Learning curves were computed for the best performing number of nodes for early, intermediate and late fusion in data set A in a "leave two out" cross validation. In each fold, a complete learning curve was estimated varying

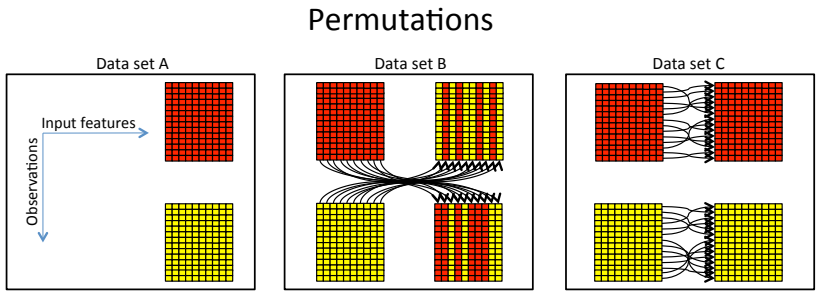

Fig. 1. The three data sets analysed. A is the original set with input features selected using a simple univariate test. In data set B the features are permuted between modalities. In data set C the observations are permuted group wise for one modality.

the size of training data from 2 to 78, but keeping group proportions in training data fixed at 50%. Both the 8 fold- and the "leave two out" cross validation experiments were repeated 300 times each.

III. Results and Discussion

Generalization performance as assessed by crossvalidation for the different combinations of early, intermediate, and late integration experiments on data set A, B, and C, are shown in table I. The overall best performance is obtained for late integration when based on data sets A or C. For data set B, with presumably high inter-modal dependency, early integration shows the best performance, and is equivalent to the performance seen in early integration model on data set A ($p = 0.7$). For completeness we also tested performance of the models trained on the individual modalities separately, finding somewhat higher test errors.

The results from the grid search experiment on the individual modalities and on the three levels of fusion of data set A are shown in Fig. 3. The red square marks the overall lowest error, which is seen in late fusion with one node for the functional modality and two for the structural. For the early fusion, a peak in error is seen for 2 nodes. An increase at two nodes is also seen for the individual modalities, which could be an indication that in this case, the RBM finds some alternative hidden variables e.g. age or gender. For the late and especially for the intermediate fusion, an increase in error is seen as more nodes are included. The lowest error is in both cases found centred around (2,2) nodes.

Learning curves for early, intermediate, and late data fusion of data set A are seen in Fig. 4. From this it is seen that the hidden unit selection scheme seems to prevent overfitting.

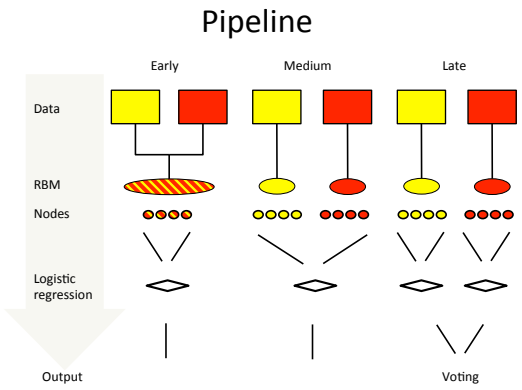

Fig. 2. Early, intermediate, and late fusion pipelines.

TABLE I
Node combination, mean error, and standard error of the mean from individual modalities and for early (E), intermediate (I), and late (L) fusion of data sets A, B, and C. Lowest error is bold for each data set. For E, I, and L, the p-value from a two-sample t-test between the fusion levels are shown. As the comparison is commutative, only the lower triangle of the 3x3 matrix is shown.

Individual modalities	No Nodes F	No Nodes S	Err	SE		
Functional	1		2.101e-01	2.517e-04		
Structural	1		2.752e-01	5.755e-04		
Data set A	**No Nodes F**	**No Nodes S**	**Err**	**SE**	**Stat Diff (p) E**	**Stat Diff (p) I**
Early	3		1.785e-01	1.346e-03	-	-
Intermediate	1	2	1.907e-01	1.312e-03	2e-10	-
Late	1	2	**1.748e-01**	1.122e-03	4e-02	6e-19
Data set B	**No Nodes F/S**	**No Nodes F/S**	**Err**	**SE**	**Stat Diff (p) E**	**Stat Diff (p) I**
Early	3		**1.793e-01**	1.444e-03	-	-
Intermediate	3	3	2.092e-01	1.805e-03	8e-34	-
Late	4	4	2.001e-01	1.615e-03	2e-20	2e-04
Data set C	**No Nodes F**	**No Nodes S**	**Err**	**SE**	**Stat Diff (p) E**	**Stat Diff (p) I**
Early	1		1.834e-01	1.117e-03	-	-
Intermediate	1	2	1.843e-01	1.294e-03	6e-01	-
Late	1	2	**1.752e-01**	1.131e-03	4e-07	1e-07

Rather low dimensional models are chosen, except in the data set B. The grid search results in Fig 3 illustrates that the model is inclined to choosing a simple model.

The results on the original data and the two permuted data sets together present evidence for the basic conditional independence hypothesis: The label patient/control is the strongest link between the two modalities. The primary evidence is the statistically significant (at 5% significance level) improved performance of the late fusion model in data set A. In addition, the fact that the late fusion model on the permuted set C achieves the lowest performance of the three fusion levels (and the same performance as seen in A) supports this hypothesis. The permutation scheme breaks any other dependency structure between the two modalities, though it still maintains performance, and with a model of same complexity (number of hidden RBM units).

The fact that data set B, which presumably has strong inter-modality dependencies, shows poor performance when modelled with intermediate and late fusion architectures, again is evidence that with the given signal-to-noise ratio and sample sizes we are in fact able to detect dependencies, when they exist.

IV. Conclusion

Our results provide evidence that the proposed permutation strategies can elicit the conditional dependency structure among modalities in a multimodal decision problem. The present work is to the best of our knowledge the first to use such permutation schemes. Future work should be conducted to expand and test the procedures on a broader selection of data sets, and also to further investigate the effects of the feature and model selections. The implications for Schizophrenia diagnosis support should likewise be explored to a greater extend.

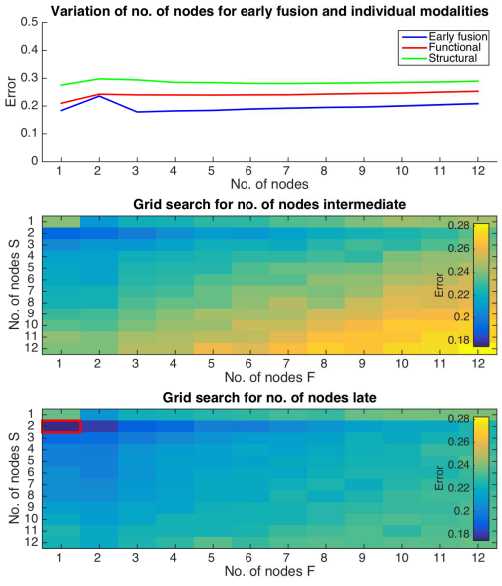

Fig. 3. Grid search for optimal number of nodes for early, intermediate, and late integration of the two modalities, fMRI (F) and sMRI (S), of data set A. A. The red square denotes the best performing node combination in the best performing fusion level (late, F=1, S=2).

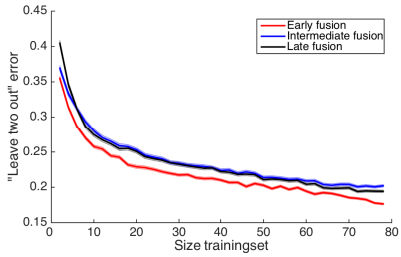

Fig. 4. Learning curve for early, intermediate and late fusion of data set A. The best performing node combinations for each fusion level were analysed for the learning curve. These were Early: 3, Intermediate: (1,2), Late: (1,2).

Acknowledgment

We gratefully acknowledge the data shared by the Mind Research Network, and funded by a Center of Biomedical Research Excellence (COBRE) grant 5P20RR021938/P20GM103472 from the NIH to Dr. Vince Calhoun. The current study was supported by the Lundbeck Foundation (R155-2013-16337).

References

[1] M. M. Picchioni and R. M. Murray, "Schizophrenia," *BMJ*, vol. 335, no. 7610, pp. 91–95, 2007.

[2] D. D. L. Hall and J. Llinas, "An introduction to multisensor data fusion," *Proceedings of the IEEE*, vol. 85, no. 1, pp. 6–23, 1997.

[3] C. Pohl and J. L. Van Genderen, "Review article Multisensor image fusion in remote sensing: Concepts, methods and applications," *International Journal of Remote Sensing*, vol. 19, pp. 823–854, 1998.

[4] P. K. Atrey, M. A. Hossain, A. El Saddik, and M. S. Kankanhalli, "Multimodal fusion for multimedia analysis: A survey," *Multimedia Systems*, vol. 16, pp. 345–379, 2010.

[5] J. Sui, T. Adali, Q. Yu, J. Chen, and V. D. Calhoun, "A review of multivariate methods for multimodal fusion of brain imaging data," *Journal of Neuroscience Methods*, vol. 204, no. 1, pp. 68–81, 2012.

[6] S. Xiang, L. Yuan, W. Fan, Y. Wang, P. M. Thompson, and J. Ye, "Bi-level multi-source learning for heterogeneous block-wise missing data," *NeuroImage*, vol. 102, pp. 192–206, 2013.

[7] L. Itti and P. Baldi, "Bayesian surprise attracts human attention," *Vision Research*, vol. 49, pp. 1295–1306, 2009.

[8] M. S. Çetin, F. Christensen, C. C. Abbott, J. M. Stephen, A. R. Mayer, J. M. Cañive, J. R. Bustillo, G. D. Pearlson, and V. D. Calhoun, "Thalamus and posterior temporal lobe show greater inter-network connectivity at rest and across sensory paradigms in schizophrenia," *NeuroImage*, vol. 97, pp. 117–126, 2014.

[9] R. Kohavi and G. H. John, "Wrappers for feature subset selection," *Artificial Intelligence*, vol. 97, no. 97, pp. 273–324, 1997.

[10] R. B. Palm, "Prediction as a candidate for learning deep hierarchical models of data," Technical University of Denmark, IMM, Tech. Rep., 2012. [Online]. Available: https://github.com/rasmusbergpalm/DeepLearnToolbox

[11] G. E. Hinton, "Training products of experts by minimizing contrastive divergence." *Neural computation*, vol. 14, pp. 1771–1800, 2002.

[12] ——, "A Practical Guide to Training Restricted Boltzmann Machines A Practical Guide to Training Restricted Boltzmann Machines," *Computer*, vol. 9, p. 1, 2010.

Appendix B

Paper B

Transferability of learning efficiency from quick learning generative models

---

Unpublished preprint

[Axelsen, M. C., Bak, N., and Hansen, L. K. 2019. "Transferability of learning efficiency from quick learning generative models." - To be submitted]

Transferability of learning efficiency from quick learning generative models

Martin C.. Axelsen[*]
Section for Cognitive Systems
DTU Compute
Technical University of Denmark
2800 Kgs. Lyngby, Denmark.
maxe [at] dtu.dk

Nikolaj Bak[†]
CINS and CNSR[‡]
Mental Health Services Glostrup
Capital Region, Denmark.

Lars Kai Hansen
Section for Cognitive Systems
DTU Compute
Technical University of Denmark
2800 Kgs. Lyngby, Denmark.

Abstract

High classification performance can be a challenge when working with small sample sizes, and model choices can significantly influence performance. It is well known that some generative models are quick learners, i.e., need less data to achieve asymptotic performance, compared to similar complexity discriminative models. As synthetic data can be created by generative models, additional training data can be fed to the slower learning model in an attempt to improve its learning efficiency. To examine the feasibility of this data augmentation approach, we investigate the generative/discriminative model pair naïve Bayes and logistic regression. The models are initially evaluated on synthetic data of different distributions and with different assumptions of covariances. We find that the augmentation method does indeed work, when the assumptions of the generative model are close to being fulfilled. Subsequently eight datasets from the UCI database are evaluated with the two models in a learning curve manner. We find that in most cases, the data augmentation approach improves classification performance beyond the performance of the naïve Bayes and logistic regression models used for comparison.

Introduction

In many research fields, e.g. neuropsychiatry and neuroscience, obtaining large amounts of data is difficult and/or expensive. Machine learning methods are increasingly used for analyzing data from these domains. In machine learning, some classification models tend to learn quickly, while others require more training data, even within models with comparable complexity, making model choice critically dependent on sample size [21]. For smaller amounts of data, data-efficient quick learning methods are useful.

So-called generative models for classification aim at modeling the whole label-feature distribution, hence solving an intrinsically more complex problem than so-called discriminative classification

[*]Corresponding author.
[†]Now works at H. Lundbeck A/S
[‡]CINS - Center for Clinical Intervention and Neuropsychiatric Schizophrenia Research and CNSR - the Center for Neuropsychiatric Schizophrenia Research

models. Interestingly, several studies have found that generative models can be quicker learners compared to discriminative models [21, 23].

To test how quick a model learns, and to compare between models, learning curves can be used [24]. In this context, a learning curve is a plot of model performance, e.g. classification error, as a function of training set size. The learning curve of a quick learning model will be steep initially, and will reach an asymptotic level with few training samples. A slower learning model will need more training samples to reach an asymptotic level, however the asymptotic performance will often be better compared to the quicker learning model [21], meaning that their learning curves will intersect [19].

When set with a classification task based on a large data set, and where good discriminative performance is the primary goal, a slow learning model will often be the better performing model and thereby the obvious choice. When operating with smaller amounts of data, it is not clear if sufficient data is available for a slow learning model to outperform a quick learning model. The model choice is thus not obvious. A perfect combination of a quick and a slow learning model would yield a so-called oracle model, which at all training data sizes would share performance with the best performing model. This oracle model, though being computationally impractical, would have a learning curve that followed the lower envelope of the quick and the slow learning model.

We hypothesize that the quicker learning ability of the generative model can be transferred to a slower learning model via data augmentation. More specifically, samples generated by the generative model can be used to augment the training data of the slower learning model. In the current study, we investigate whether the above hypothesis is valid and under which conditions.

Generative models learn the joint probability $p(x, y)$ of data x and label y from training data. Through Bayes' rule, the joint probability can be used for classification using the posterior probability distribution $p(y|x)$, which according to Bayes' decision theorem minimizes the classification error rate and hence should be optimal

$$p(x, y) = p(x|y)p(y) = p(y|x)p(x) \Rightarrow$$
$$p(y|x) = \frac{p(x|y)p(y)}{p(x)}.$$

As the joint probability of x and y, $p(x, y)$ is modeled, it is possible to sample from this distribution and thereby generate new labeled synthetic samples - hence the name generative models. Discriminative approaches model the conditional distribution $p(y|x)$ directly.

The interplay and combination of generative and discriminative models have been investigated in several studies, (see e.g. [21, 23, 27, 17, 22]). In the study by Perina and colleagues [22], methods for combining generative and discriminative models are placed in three categories: Blending, Iterative, and Staged methods. Blending methods, also called hybrid learning, is where a hybrid objective function containing at least a discriminative and a generative term is optimized. In iterative methods, generative and discriminative models, that complement each other, when learning are trained iteratively. Finally, in staged methods, a generative model learns a set of features from the data, and a discriminative classifier is then trained on the derived features.

The analysis of applying a generative model to generate data for augmenting the training set of a discriminative model shares similarities with the line of thought in staged methods.

Data augmentation is widely used to fit a slower learning model (or a more complex model) to a problem where an insufficient number of labeled samples are available and can be performed in either a latent feature space or directly in data space [4].

Data augmentation has been used for over two decades in classification of image content, where manual labeling of the data is a bottleneck [26]. Especially for fitting convolutional neural networks (CNN), data augmentation is essential to avoid overfitting of the model [15, 14]. To ensure that the label of the synthetic data is reasonable, data augmentation in data space is most often used in this domain. Typical data space augmentation schemes include various degrees of rotations, scalings, translations and warping of the images. Many of these schemes have been developed for recognition of handwritten digits or letters, where rotation especially can be an issue. In the work by Ha and Bunke [11], an augmentation scheme was developed to imitate other types of handwriting and other types of writing instruments through minor rotations, skewing and erosion and dilation of the pixels. A more recent method of feature space data augmentation was proposed by DeVries and colleagues

[6], who use a sequential auto-encoder to learn a latent space and subsequently interpolate and extrapolate between observations in the latent space to generate synthetic data similarly to [4].

In generative adversarial nets (GANs) [9], a generative and a discriminative model are combined in order to estimate a potentially intractable data distribution. This is done by training a generative model to create adversarial data examples that the discriminative model subsequently tries to distinguish from true examples. The generative model then learns how to generate examples that are indistinguishable from the true data to the discriminator, and thereby estimate the true generative distribution. Recently, Antoniou and colleagues [1] expanded the GAN framework to employ the estimated distribution to create synthetic examples for data augmentation. They evaluated the method on two handwriting data sets and a face recognition data set.

In the field of learning from imbalanced data sets, generation of synthetic data in both data and feature space has been investigated. Chawla and colleagues [4] proposed the SMOTE algorithm, which combines an oversampling of the minority class with generation of synthetic samples through a nearest neighbor interpolation scheme. This was further expanded by Han and colleagues [12] to only generate synthetic samples close to the decision boarder. In work by Guo and Viktor [10], the Databoost method for combining a boosting algorithm with data generation is proposed for imbalanced data sets. More specifically, hard-to-classify examples are identified for both the majority and minority class via the boosting algorithm. These examples are then used to generate synthetic examples in feature space, which the training set is then augmented with. For continuous variables, the synthetic examples are generated via a naïve Bayes method by sampling from normal distributions with mean and variance found empirically from the identified hard examples. In work by Shaikhina and Khovanova [25] surrogate data is created for a regression task by sampling independently from normal distributions with empirical mean and standard deviations based on original features. The relationship between the input data and the variable they seek to predict is thus disregarded.

Here, we expand on these methods and investigate them in the context of the learning efficiency. Furthermore we explore the limits of such an approach with a specific focus on the importance of sample size for the data augmentation step. In the studies by Ng and Jordan [21], and Xue and Titterington [27], the asymptotic performance of a generative (naïve Bayes) and a discriminative (logistic regression) model is compared using learning curves. The naïve Bayes/logistic regression model pair is part of the same parametric family [21] and has been shown to exhibit an example of a quick and a slow learning model, respectively [21, 27]. In the current study, we make use of the same two models and a similar model comparison framework to validate the possibility of transferring the quick learning ability of the generative naïve Bayes to the slower learning discriminative logistic regression model via data augmentation. A logistic regression model is trained on augmented training data and compared with the predictive performance of a naïve Bayes and a logistic regression model trained on un-augmented training data. Learning curves are used to compare the models across data sizes.

1 Methods

1.1 Models

The naïve Bayes model used in the current study is based on the normal distribution with separate means for each class, but with a common pooled covariance, similar to what was used by Ng and Jordan [21]. This version of naïve Bayes is equivalent to a normal-based linear discriminant analysis (LDA) model with a diagonal covariance matrix [27][13, p. 108]. For binary classification, the posterior is evaluated as

$$\begin{aligned} p(y=1|x) &= \frac{f_1(x)\cdot\pi_1}{f_1(x)\cdot\pi_1 + f_2(x)\cdot\pi_2} \\ &= \sigma\left(\log\left(\frac{f_1(x)}{f_2(x)}\right) + \log\left(\frac{\pi_1}{\pi_2}\right)\right), \end{aligned}$$

where $f_k(x)$, $k = \{1, 2\}$ is the class conditional likelihood, π_k is the class conditional prior, and σ is the sigmoid function ($\sigma(x) = (1 + \exp(-x))^{-1}$). This can be shown to be identical to a logistic regression model [13, p. 127]

$$p(y|x) = \sigma(\alpha_0 + \alpha^T x),$$

where, if $f_k(x)$ is normal, then follows

$$\alpha_0 = \log\left(\frac{\pi_1}{\pi_2}\right) - \frac{1}{2}(\mu_1 + \mu_2)\Lambda^{-1}(\mu_1 - \mu_2)^T$$
$$\alpha = \Lambda^{-1}(\mu_1 - \mu_2)^T$$

where $\Lambda = \text{diag}(\sigma_1^2, \ldots, \sigma_D^2)$, σ_d^2 is the variance of the d'th variable and μ_k is the mean vector of the k'th class, and D is the dimensionality.

As done by Ng and Jordan in [21] in order to avoid division by zero, we apply a small regularization to the model by adding $\epsilon = 2^{-52}$ (i.e. the spacing of double precision floating point numbers) to the variances.

For naïve Bayes to learn a model of the data, the empirical class conditional mean $\hat{\mu}_k$ and the empirical pooled diagonal covariance matrix $\hat{\Lambda}$ is calculated from training data.

$$\hat{\sigma}_d^2 = \frac{\sum_k m_k \cdot \hat{\sigma}_{d,k}^2}{m} \Rightarrow \tag{1}$$
$$\hat{\Lambda} = \text{diag}(\hat{\sigma}_1^2, \ldots, \hat{\sigma}_D^2), \tag{2}$$

where m is the number of data samples in the training set, m_k is the number of data samples in each class (k) of the training set ($m = \sum_k m_k$), and $\hat{\sigma}_{d,k}^2$ is the empirical variance of class k in feature d.

Synthetic data is generated from the naïve Bayes model by sampling separately for each class from a multivariate normal distribution with the empirical class conditional mean vector and diagonal covariance matrix

$$\hat{x}_k \sim \mathcal{N}(\hat{\mu}_k, \hat{\Lambda}). \tag{3}$$

1.1.1 Learning curve analysis framework

For a given training set, a learning curve is constructed by incrementally increasing the amount of training data available to the model. Each trained model is subsequently evaluated on the held out test data.

For each increment (i) along the learning curve, and for each data set, a naïve Bayes model and a logistic regression model were fitted to the available training data (x_i), with number of samples m_i. Furthermore, a logistic regression model was fitted to the same data (x_i), augmented with synthetic data. Synthetic data for each class ($\hat{x}_{i,k}$) was generated according to equation 3 based on the class conditional empirical mean and empirical pooled diagonal covariance matrix of the available training data

$$\hat{x}_{i,k} \sim \mathcal{N}(\hat{\mu}_{i,k}, \hat{\Lambda}_i).$$

All learning curve analyses were done stratified, meaning that the class proportions were constant for all analyses within a given data set.

1.2 Simulated data

To investigate the quick and slow learning properties of the naïve Bayes model and the logistic regression model and to look into the properties of the synthetic data generated from the naïve Bayes model, data was simulated following the same approach as in [27]. Four different data generating distributions were used and for each of these, four different covariance matrices were applied, see table 1. Data was simulated having two dimensions, with the following distributions and covariance matrices:

- Normal distribution, $X \sim N(\mu_k, \Sigma_k)$
- Student t-distribution with degrees of freedom $\nu = 3$
- Log normal distribution $ln(X) \sim N(\mu_k, \Sigma_k)$
- Mixed normal distribution $X \sim N(\mu_{(k,l)}, \Sigma_k)$

Table 1: **Covariance matrices and means of the simulated data.**

Shared diagonal covariance	Shared full covariance
$\Sigma_1 = \Sigma_2 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$ $\mu_1 = [1\ ,\ 0]\ ,\ \mu_2 = [-1\ ,\ 0]$	$\Sigma_1 = \Sigma_2 = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}$ $\mu_1 = [1\ ,\ 0]\ ,\ \mu_2 = [-1\ ,\ 0]$
Separate diagonal covariances	**Separate full covariances**
$\Sigma_1 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \Sigma_2 = \begin{bmatrix} 0.25 & 0 \\ 0 & 0.75 \end{bmatrix}$ $\mu_1 = [1\ ,\ 0]\ ,\ \mu_2 = [-1\ ,\ 0]$	$\Sigma_1 = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}, \Sigma_2 = \begin{bmatrix} 0.25 & 0.5 \\ 0.5 & 1.75 \end{bmatrix}$ $\mu_{(1,1)} = [1\ ,\ 0]\ ,\ \mu_{(2,1)} = [-1\ ,\ 0]$ $\mu_{(1,2)} = [3\ ,\ 0]\ ,\ \mu_{(2,2)} = [-3\ ,\ 0]$

Where $k = \{1, 2\}$ indicates class and $l = \{1, 2\}$ indicates latent class in the mixed normal distribution.

From each distribution and covariance combination, 1000 samples were simulated with equal class distributions.

1.2.1 Analysis pipeline for simulated data sets

For each of the 16 conditions, data sets were randomly simulated 200 times. Learning curve experiments were done on a random stratified split of 20% training data and evaluated on the remaining 80% test data of each simulated data set. For each increment along the learning curve, a naïve Bayes model and a logistic regression model were trained on the available training data (x_i). Furthermore a logistic regression model was trained on the available training data augmented with $gN = [50, 100, 200, 400, 800]$ synthetic samples, with same class distribution as the true training data, i.e. equal classes. All models from all learning curve increments were evaluated on the test set.

1.3 UCI data sets

Eight data sets from the UCI repository [7] were used in the current study. These eight data sets were also investigated in [21] and [27]. For some data sets, continuous, binary, and categorical variables were available. In these cases, following [21] and [27], only the continuous and binary variables were included. The data sets span different domains, from medical data, across financial and different scientific measurements to image data, and were chosen to test the diversity and robustness of the proposed approach. Variables with zero variance across all samples were removed from each data set. Table 2 shows the details of the individual data sets.

Table 2: **Data sets used for evaluating the data augmentation approach.** For each data set, the number of samples (N), the number of samples in test set ($N_{20\%}$), the proportion of classes in dataset (Class balance), the number of dimensions (D), and the number of dimensions with non-zero variance (D_v) are shown.

data set	N	Test size ($N_{20\%}$)	Class balance	D	Dv
Pima	768	154	0.35	8	8
Adult	48842	9768	0.24	6	6
Boston	506	101	0.49	13	13
Optdigits 0 vs. 1	765	153	0.49	64	52
Optdigits 2 vs. 3	769	154	0.49	64	55
Ionosphere	351	70	0.36	34	33
Sonar	208	42	0.47	60	60
Liver disorders	341	68	0.42	6	6

1.3.1 Analysis pipeline for UCI data sets

All models were trained with 1000 random splits of 80% training data and 20% test data. Stratified learning curves were created from training data. All models from all learning curve increments were

evaluated on the test set. The learning curves were initiated at a data size of $m_0 = Dv + \Delta$ so as to avoid underdetermined models.

1.3.2 Model comparison for the UCI data sets

The logistic regression model based on augmented data (the augmented model) was compared with the naïve Bayes model for small amounts of training data, i.e. up until the learning curves of the logistic regression and naïve Bayes models cross (at m_c) (see figure 1). We thus compare the models for a range of data set sizes where the slower learning model will eventually be superior as the quicker learning model will be close to reaching an asymptotic error. Experiments with a range of numbers of generated samples $gN = [50, 100, \ldots, 950, 1000]$ were conducted to evaluate the influence of this on the performance of the augmented model.

For each data set, the models are compared at the closest-measured learning curve increment after m_c (denoted m_{c+}). The models are compared using the Bayesian analysis framework of Benavoli and colleagues in [2] using the Bayesian correlated t-test [5], which makes a paired comparison of the errors from each of the 1000 random splits of training and test data. The Bayesian correlated t-test estimates the mean difference of accuracy between the two models being compared, while taking the correlation due to overlapping training and test sets into account using the heuristic suggested by Nadeau and Bengio [20]. From the Bayesian correlated t-test, the posterior of the mean difference of accuracy is obtained and evaluated in relation to a so-called region of practical equivalence (ROPE) [16]. When 95% of the posterior probability is on either side or within the ROPE, the result is conclusive, i.e. one model has higher performance or their performances are equal, respectively. Otherwise, the test result is deemed inconclusive [2, 16] and hence needs further investigation. The ROPE is set to 0.01 as recommended for comparison of classifiers by Benavoli and colleagues [2]. Publicly available Python code was used for the comparison [2].

To further investigate the behavior of the augmented model, the sample efficiency is calculated within the small data segment of the learning curves. The sample efficiency is defined as the difference in number of true samples that is required by the oracle model to achieve a similar mean performance

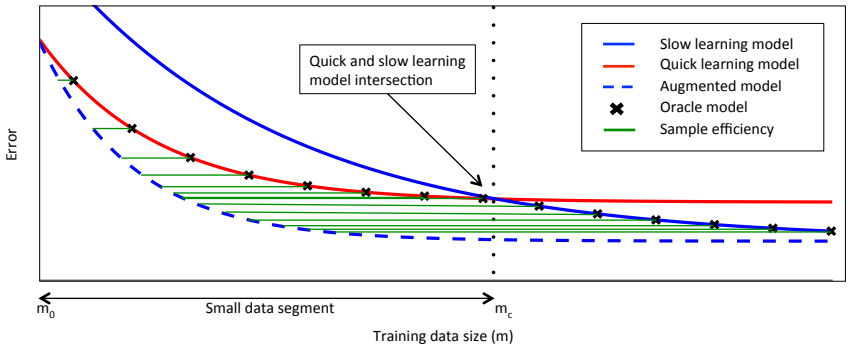

Figure 1: **Illustration of the learning curve method.** The appearance of learning curves of a quick (red) and a slower (blue) learning model, the combination of these, i.e. the oracle model (black crosses) and the theoretical appearance of an augmented model (dashed blue). The crossing of the learning curves of the quick and the slow learning model (m_c) is denoted to illustrate the small data segment. Furthermore, the sample efficiency of the augmented model is illustrated with the green vertical lines indicating the difference in number of true samples used for training the augmented model and the oracle model with equal performance.

as the augmented model relative to the number of samples needed by the augmented model (see equation 4 and figure 1).

$$\text{Sample efficiency } (\%) = \frac{N_{tr}^{OR} - N_{tr}^{AM}}{N_{tr}^{AM}} \cdot 100, \tag{4}$$

where N_{tr}^{OR} is the number of true training samples needed by the oracle model to reach an equivalent performance to the augmented model trained on N_{tr}^{AM} true data samples.

To investigate the behavior of the augmented model when more training data is available, the learning curves are considered with a focus on their asymptotic behavior.

Apart from the statistical comparison with the Bayesian analysis, all processing was performed using Matlab [18]. The standard implementation of logistic regression in Matlab called `glmfit` was used. The naïve Bayes model for classification and generation of data was implemented in Matlab and is available from Github: `https://github.com/mcaxelsen/gm_augmentation`.

2 Results

2.1 Results from simulated data

From the plots in figure 2 it is seen that for the log-normal distributed data the naïve Bayes model reaches an asymptotic performance very fast, i.e. with a very small amount of training data. However, the performance obtained is much worse compared to the performance of the logistic regression model, which seems to learn equally fast for these datasets. For the distributions where the datasets have a diagonal covariance matrix, the naïve Bayes model is the quicker learner. In the full covariance cases, the logistic regression model is the quicker learner for all distributions.

It is seen that the logistic regression model augmented with synthetic data in most situations lies between the performance of the baseline models, and that increasing the amount of synthetic data used for augmentation brings the performance closer to the naïve Bayes model and vice versa. Interestingly, for the data simulated from a normal distribution with separate diagonal covariance matrices per class, the augmented model outperforms the two baseline models for some numbers of generated samples.

2.2 Results from UCI data

The data augmentation approach is further investigated through experiments on eight data sets from the UCI repository (see table 2). Initially, experiments are analyzed with a perspective on small amounts of available training data. Afterwards, more training data is added and the asymptotic behavior of the approach is investigated.

2.2.1 Small amounts of training data

In the present section, results from the comparison of the models trained on smaller amounts of training data are shown.

Using the Bayesian analysis framework, as described in section 1.3.2, the oracle model and the augmented model are compared. The point of comparison is the closest-measured learning curve increment of the naïve Bayes model and the logistic regression model after they cross (i.e. the number of samples used for training is m_{c+}). The results based on the analyses are shown in figure 3.

For each of the eight data sets, twenty augmentation sizes, from 50 to 1000 synthetic samples, are investigated, yielding a total of 160 experiments. In 121 out of 160 experiments, the augmented model is either equal to the oracle model or better (lighter blue colors in figure 3). However, 39 out of 160 experiments come out inconclusive (dark blue in figure 3), and these experiments hence need further investigation. Though we cannot apply the classification system directly in these cases, we can visualize the uncertainty in the estimation with plots of the posteriors. For all experiments on the Liver disorders data set, the posterior mass is located across the ROPE, indicating that a definitive conclusion as to which model is performing better cannot be made using this method (the posteriors for the inconclusive experiment are shown in section 4 in the supplementary material). It should be noted that for this data set the intersection point (m_c) coincides with the starting point of the learning

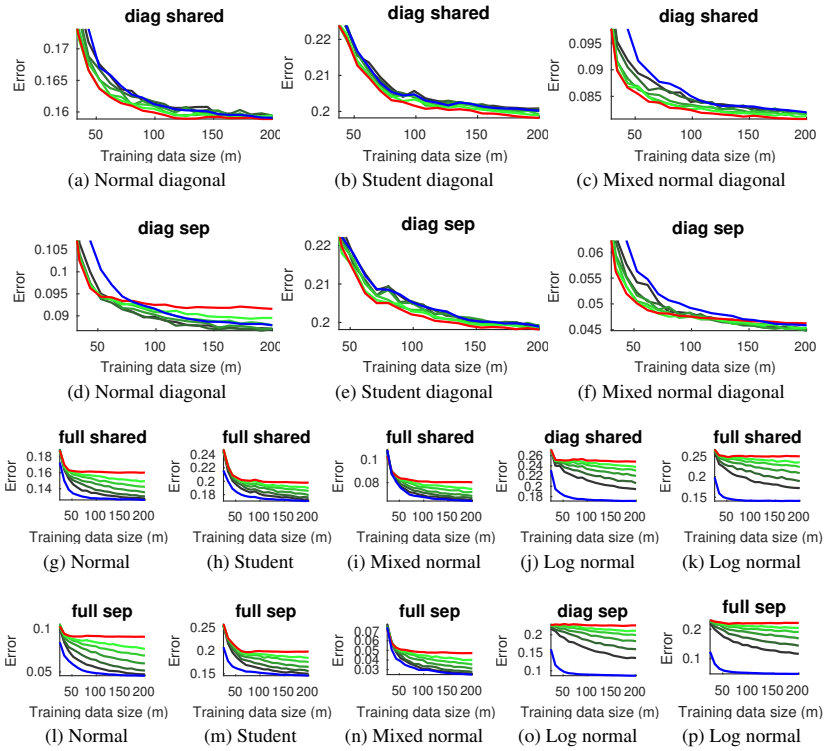

Figure 2: **Learning curves of analyses on the simulated data which show quick learning property of naïve Bayes.** Data distributions are normal distribution (a,d), student t-distribution (b,d), and mixed normal distribution (c,f) with parameters shown in table 1. The naïve Bayes model is in red, the logistic regression model is in blue, and the augmented models, calculated for 50, 100, 200, 400, and 800 synthetic samples used for augmentation, are plotted with an increasing brightness in green to illustrate that more samples are used.

curve (m_0). For the remaining data sets when looking into the inconclusive experiments, the posterior mass is located such that less than 5% of the posterior mass is towards the oracle model in relation to the ROPE. This indicates that the augmented model is equal to or better than the oracle model at the intersection point m_{c+} for all the remaining seven data sets for all experiments.

The sample efficiency as described in section 1.3.2 is now used to compare the models. The min-max ranges for the small data segments are shown in table 3 for each data set and for each number of generated samples used for augmentation. As the intersection point (m_c) for the Liver disorders data set coincides with the starting point of the learning curve (m_0), there are no results from this data set in this analysis. However, this data set is further analyzed and discussed in the following sections. The analyses of the Sonar data set for augmentation sizes of 100 samples and higher achieve an initial mean error which is below the error of the oracle model at all points on the learning curve, hence no quantitative sample efficiency results exist for these, though the improvement is at least of 168%. From the remaining results, it can be seen that sample efficiencies well above 200% is achievable, i.e. the oracle model needs three times as much training data to perform as well as the augmented model. Data sets Pima and Adult, however, show a smaller effect, with the augmented model performing in the range of the oracle model (table 3). For data sets Sonar, Boston, Optdigits 0 vs. 1 and 2 vs. 3, and Ionosphere, substantial benefits from data augmentation are evident.

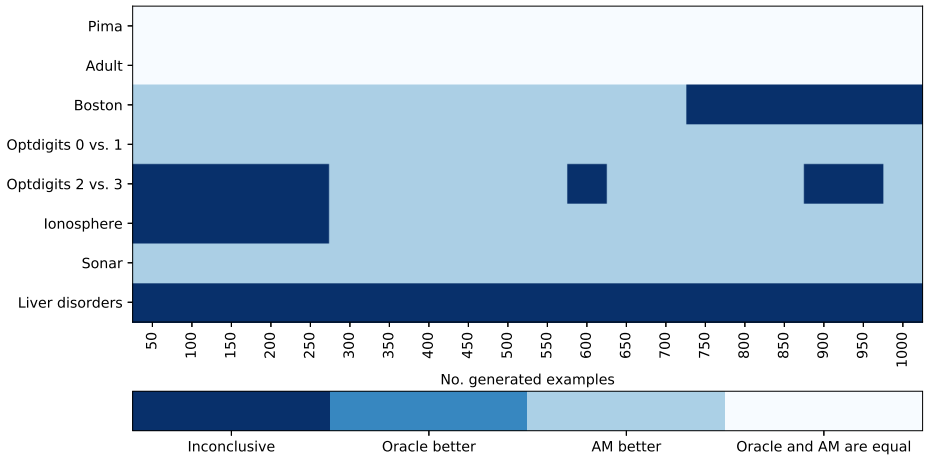

Figure 3: **ROPE analysis.** Illustration of the decisions based on the location of 95% of the posterior mass in relation to the region of practical equivalence (ROPE). AM represents the augmented model.

2.2.2 Asymptotic behavior

The learning curves from four of the eight UCI data sets are displayed in figure 4 (the remaining are shown in appendix S2 in the supplementary material). The four selected data sets each illustrates a different behavior of the trained models. The learning curves of the Pima data set (in figure 4A) and the Adult data set (in supplementary) have similar behavior. Here the augmented model acts as hypothesized when the amount of training data is less than where the two base models intersect; the data augmentation scheme brings the performance of the augmented model to the level of the oracle model i.e. the naïve Bayes model, and a high amount of generated samples seems to be better. After the intersection, the generated samples decrease the performance of the augmented model relative to the un-augmented logistic regression model. This is especially true for the Pima data set. The asymptotic behavior of the augmented model for the Pima data set seems to be parallel to the oracle model, only with a slightly higher error. For the Adult data set, the performance of the augmented model is indistinguishable from the oracle model when training on all training data, though this data set is much larger than the others (see table 2).

The results from the Liver disorders data set in figure 4D are similar in that for increasing augmentation, the performance of the augmented model approach that of the naïve Bayes model. The difference is that at no training data size is the naïve Bayes model performing better than the logistic regression model, and neither is the augmented model.

For the data sets "Boston" and "Ionosphere" shown in figures 4B and 4C, it is seen that there is an increased effect from the data augmentation at smaller amounts of training data for most amounts of generated samples. The effect is more pronounced before the intersection, however for some amounts of generated samples, the effect persists even after the intersection. For the Boston dataset (figure 4B), the trend that more generated samples bring the augmented model closer to the naïve Bayes model is seen again. However, when only augmenting the training data with smaller amounts of synthetic samples, there seems to be an overall improvement in the augmented model compared to the oracle model. The learning curve for the model augmented with 200 synthetic samples reaches the level of the oracle model, despite showing worse performance for a range of training data sizes (approximately $m = 100 \ldots 200+$). For the Ionosphere data set (and Optsdigits 0 vs 1, 2 vs 3, and Sonar), it seems that more synthetic samples improves the performance of the augmented model, which is opposite to what is seen in the other data sets. The effect does, however, seem to saturate, i.e. the performances of augmenting with 400 and 800 generated samples are similar (figure 4C). In table 3, it is similarly seen that the sample efficiency ranges are very similar for experiments with generated samples above 350. For the Sonar data set, augmenting the training data with 400 synthetic

Table 3: **Table of minimum and maximum sample efficiency within the small data segment.**When the augmented model is initially performing better than the asymptotic performance of the oracle model, the sample efficiency is denoted as being better than the sample efficiency calculated based on the full training data size (N). Positive values indicate that the augmented model outperforms the oracle model.

gN	Pima	Adult	Boston	Optdigits 0 vs. 1	Optdigits 2 vs. 3	Ionosphere	Sonar	Liver disorders
50	-31% – 1%	-18% – 15%	9% – 149%	105% – 131%	48% – 79%	-33% – 33%	58% – 98%	-
100	-22% – 9%	0% – 24%	53% – 189%	156% – 180%	76% – 114%	-2% – 70%	$> 168\%$	-
150	-20% – 4%	4% – 28%	80% – 205%	166% – 191%	78% – 131%	20% – 115%	$> 168\%$	-
200	-17% – 5%	7% – 31%	63% – 202%	168% – 207%	88% – 142%	28% – 155%	$> 168\%$	-
250	-15% – 2%	7% – 35%	60% – 212%	182% – 207%	108% – 146%	42% – 186%	$> 168\%$	-
300	-13% – 4%	11% – 32%	52% – 206%	174% – 209%	117% – 154%	53% – 201%	$> 168\%$	-
350	-9% – 4%	11% – 33%	44% – 209%	174% – 211%	124% – 154%	55% – 215%	$> 168\%$	-
400	-9% – 2%	11% – 33%	39% – 194%	182% – 208%	105% – 166%	54% – 227%	$> 168\%$	-
450	-7% – 1%	14% – 33%	36% – 195%	171% – 211%	119% – 146%	60% – 219%	$> 168\%$	-
500	-6% – 5%	14% – 34%	34% – 189%	186% – 208%	118% – 172%	64% – 227%	$> 168\%$	-
550	-10% – 6%	14% – 38%	33% – 194%	175% – 210%	108% – 166%	65% – 232%	$> 168\%$	-
600	-7% – 6%	14% – 37%	30% – 190%	181% – 205%	119% – 153%	69% – 239%	$> 168\%$	-
650	-5% – 7%	14% – 40%	27% – 190%	171% – 208%	117% – 144%	64% – 234%	$> 168\%$	-
700	-5% – 5%	14% – 36%	26% – 187%	183% – 208%	108% – 146%	69% – 234%	$> 168\%$	-
750	-11% – 4%	17% – 35%	23% – 184%	175% – 209%	94% – 144%	70% – 228%	$> 168\%$	-
800	-6% – 3%	14% – 35%	22% – 185%	171% – 206%	99% – 151%	69% – 220%	$> 168\%$	-
850	-9% – 2%	14% – 37%	22% – 185%	171% – 209%	115% – 142%	63% – 215%	$> 168\%$	-
900	-6% – 4%	16% – 35%	22% – 184%	178% – 203%	102% – 141%	64% – 218%	$> 168\%$	-
950	-5% – 6%	14% – 36%	20% – 185%	171% – 207%	107% – 146%	66% – 215%	$> 168\%$	-
1000	-9% – 3%	14% – 37%	19% – 185%	176% – 208%	113% – 139%	67% – 224%	$> 168\%$	-

samples yields the best of the evaluated models. For the two Optdigits data sets, there also seem to be a saturation, though the optimal number of generated samples is not identifiable from the plots.

3 Discussion

We hypothesized that quick learning generative models could assist slower learning discriminative models. To test this hypothesis, we created simulated data in the same way as was done in [27], and found that when the assumptions of the naïve Bayes model were approximately fulfilled - that is when the covariance matrix had a diagonal structure, the naïve Bayes model was indeed the quicker learner. In these cases, it was possible to improve the performance of a logistic regression model augmented with synthetic samples generated from the naïve Bayes model. When the covariance matrix was full, the naïve Bayes model showed a similar performance as when the covariance was diagonal. The logistic regression model, however, had in these cases the possibility of utilizing the additional covariance information and was hence the quicker learner, and furthermore reached a much lower asymptotic error. As the naïve Bayes model assumes Gaussian distributed variables, the log normal data constitutes a difficult problem. The logistic regression, without any data distribution assumptions, solves the problem much better.

The experiments on the UCI data, which also reproduced the results found by Ng and Jordan [21] and Xue and Titterington [27], showed that for most evaluated data sets, the naïve Bayes model is the quicker learner compared to the logistic regression model. The only data set inconsistent with this relationship found in our study and in the previous studies [21, 27] is the Liver disorders data set. In the cases where the naïve Bayes model is indeed the quicker learning model, we found that the performance of a logistic regression model can be improved to at least reach the performance of the naïve Bayes model by augmenting the training data with synthetic data generated from the naïve Bayes model using the presented augmentation approach. Furthermore, we found that data efficiency rates above 200% were achievable when only a small amount of data was available. When sufficient training data was available such that the performance of the logistic regression model had surpassed the performance of the naïve Bayes model, an improved performance could still be achieved through the data augmentation approach for five out of the eight evaluated data sets.

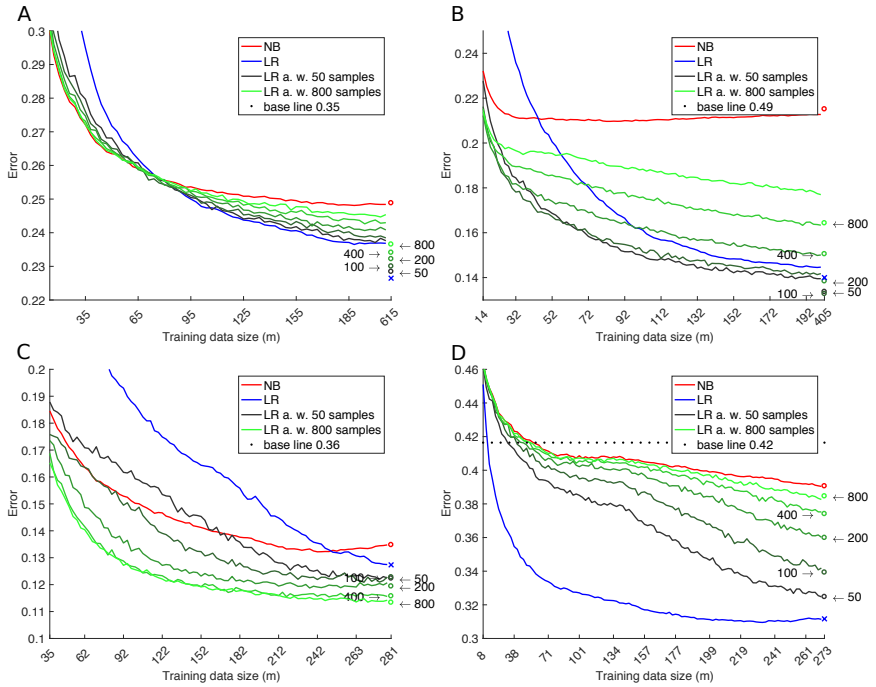

Figure 4: **Learning curves for four selected data sets.** If the data set is large, the learning curves are only calculated for training set sizes surrounding the intersection of the naïve Bayes and the logistic regression models (m_c). The colored circles at the end of the learning curves indicate the test error when training on the entire allocated training set (i.e. 80% of the data). The learning curves for the augmented models are calculated for 50, 100, 200, 400, and 800 samples used for augmentation, and are plotted with an increasing brightness in green to illustrate that more samples are used. A. Pima, B. Boston, C. Ionosphere, D. Liver disorders.

Common across all data sets, simulated and from the UCI database is that for the data sets where the generative model is the quicker learner compared to the discriminative model, the data augmentation approach improves the performance of the logistic regression model. Though the simulated data suggests that this is only the case when the naïve assumptions of the naïve Bayes model are valid, several studies have found that naïve Bayes is often a superior classifier despite the assumptions being void [8, 28]. A pragmatic approach to allow for safely using the approach is to evaluate whether naïve Bayes is the quicker learner compared to logistic regression on a training set. This would also allow for a grid search for the optimal number of synthetic samples used for augmentation, given that naïve Bayes is the quicker learner.

The augmentation scheme could be used to augment training data for any subsequent classifier, though studies should be done to verify potential differences in benefit among different classifiers. Future studies should also be done to investigate potential quick learning properties of other generative models and the potential to use synthetic data generated from these for data augmentation.

A further perspective of the current data augmentation model could be to expand it with an adversarial detection or rejection sampling scheme as used in [3] and [9], so as to potentially improve the quality of the generated samples and thereby potentially improve the performance of the classifier as well.

4 Conclusion

In many fields, large data sets are difficult to obtain, hence raising the need for generally applicable data-optimizing methods in order to efficiently make inference from the smaller data sets available. In the present study, we have shown that a quick learning generative model can assist a slower learning model in reaching a better performance through a data augmentation framework. The method was evaluated on simulated data and on continuous and binary data from different domains with positive results for all data sets where the generative model indeed is the quicker learner.

Acknowledgments

The authors would like to thank the UCI machine learning repository for providing the data sets. The current study was supported by the Lundbeck Foundation (R155-2013-16337).

Supporting information

S1 Fig. Posteriors of inconclusive results.

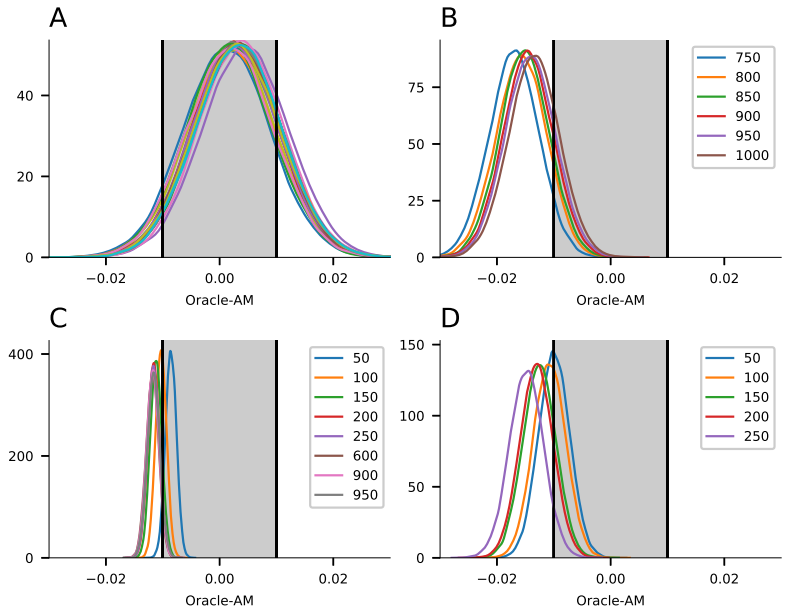

Figure 5: **Posterior probability output from inconclusive ROPE analysis.** Posterior probability for the 39 experiments with inconclusive results shown per data set. Black vertical lines and gray shade indicate the region of practical equivalence (ROPE = ± 0.01). For an experiment to be conclusive 95% of the posterior mass must be on either side or within the ROPE limit. A posterior distribution left of the left ROPE indicates that the augmented model (AM) is better, and to the right of the right ROPE indicates that the oracle model is better. If 95% of the posterior mass is within the ROPE the models are practically equivalent. Data sets are Liver disorders (A), Boston (B), Optdigits 2 vs. 3 (C), and Ionosphere (D). In Fig. 5A legend is omitted as all experiments are inconclusive.

S2 Fig. Learning curves for the remaining data sets.

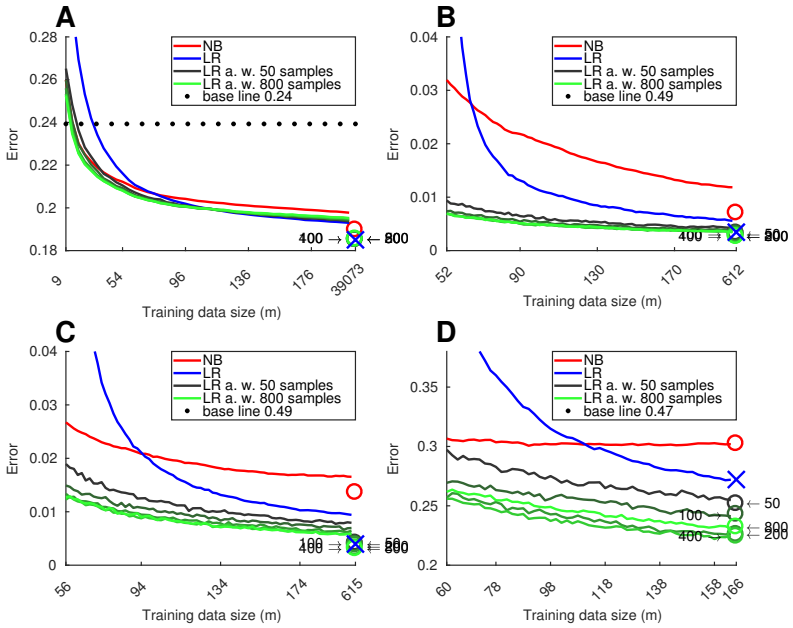

Figure 6: **Learning curves for the remaining four data sets.** If the data set is large, the learning curves are only calculated for training set sizes surrounding the intersection of the naïve Bayes and the logistic regression models (m_c). The colored circles at the end of the learning curves indicate the test error when training on the entire allocated training set (i.e. 80% of the data). The learning curves for the augmented models are calculated for 50, 100, 200, 400, and 800 samples used for augmentation, and are plotted with an increasing brightness in green to illustrate that more samples are used. A. Adult, B. Optdigits 0 vs 1, C. Optdigits 2 vs 3, D. Sonar

References

[1] Antreas Antoniou, Amos Storkey, and Harrison Edwards. Data augmentation generative adversarial networks. *arXiv preprint arXiv:1711.04340*, November 2017.

[2] Alessio Benavoli, Giorgio Corani, Janez Demšar, and Marco Zaffalon. Time for a change: a tutorial for comparing multiple classifiers through bayesian analysis. *J. Mach. Learn. Res.*, 18(77):1–36, 2017.

[3] Eduardo Castro, Jessica A. Turner, and Vince D. Calhoun. Generation of Synthetic Structural Magnetic Resonance Images for Deep Learning Pre-Training. In *IEEE 12th International Symposium on Biomedical Imaging (ISBI)*, number APRIL, pages 1057–1060, 2015.

[4] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. SMOTE: synthetic minority over-sampling technique. *J. Artif. Intell. Res.*, 16:321–357, 2002.

[5] Giorgio Corani and Alessio Benavoli. A bayesian approach for comparing cross-validated algorithms on multiple data sets. *Mach. Learn.*, 100(2-3):285–304, September 2015.

[6] Terrance DeVries and Graham W Taylor. Dataset augmentation in feature space. *arXiv [stat.ML]*, February 2017.

[7] Dua Dheeru and Efi Karra Taniskidou. UCI machine learning repository. `http://archive.ics.uci.edu/ml`, 2017.

[8] Pedro Domingos and Michael Pazzani. On the optimality of the simple bayesian classifier under Zero-One loss. *Mach. Learn.*, 29(2):103–130, November 1997.

[9] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative Adversarial Nets. *Advances in Neural Information Processing Systems 27*, pages 2672–2680, 2014.

[10] Hongyu Guo and Herna L Viktor. Learning from imbalanced data sets with boosting and data generation: The DataBoost-IM approach. *SIGKDD Explor. Newsl.*, 6(1):30–39, June 2004.

[11] T M Ha and H Bunke. Off-line, handwritten numeral recognition by perturbation method. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(5):535–539, May 1997.

[12] Hui Han, Wen-Yuan Wang, and Bing-Huan Mao. Borderline-SMOTE: A new Over-Sampling method in imbalanced data sets learning. In *Advances in Intelligent Computing*, Lecture Notes in Computer Science, pages 878–887. Springer, Berlin, Heidelberg, August 2005.

[13] Trevor Hastie, Robert J. Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*, volume 1. Springer, second edition, 2009.

[14] Søren Hauberg, Oren Freifeld, Anders Boesen Lindbo Larsen, John W. Fisher, and Lars Kai Hansen. Dreaming More Data: Class-dependent Distributions over Diffeomorphisms for Learned Data Augmentation. *Artificial Intelligence and Statistics (AISTATS)*, 51:342–350, 2016.

[15] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. ImageNet Classification with Deep Convolutional Neural Networks. In F Pereira, C J C Burges, L Bottou, and K Q Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc., 2012.

[16] John K Kruschke and Torrin M Liddell. The bayesian new statistics: Hypothesis testing, estimation, meta-analysis, and power analysis from a bayesian perspective. *Psychon. Bull. Rev.*, February 2017.

[17] Julia A. Lasserre, Christopher M. Bishop, and Thomas P. Minka. Principled hybrids of generative and discriminative models. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 1(6):87–94, 2006.

[18] The Mathworks, Inc., Natick, Massachusetts. *MATLAB version 9.1.0.441655 (R2016b)*, 2016.

[19] Niels Mørch, Lars K Hansen, Stephen C Strother, Claus Svarer, David A Rottenberg, Benny Lautrup, Robert Savoy, and Olaf B Paulson. Nonlinear versus linear models in functional neuroimaging: Learning curves and generalization crossover. In *Information Processing in Medical Imaging*, Lecture Notes in Computer Science, pages 259–270. Springer, Berlin, Heidelberg, June 1997.

[20] Claude Nadeau and Yoshua Bengio. Inference for the generalization error. *Mach. Learn.*, 52(3):239–281, September 2003.

[21] Andrew Y. Ng and Michael I. Jordan. On Discriminative vs. Generative Classifiers: A comparison of logistic regression and naive Bayes. In T G Dietterich, S Becker, and Z Ghahramani, editors, *Advances in neural information processing systems*, number 15, pages 841–848. MIT Press, 2002.

[22] Alessandro Perina, Marco Cristani, Umberto Castellani, Vittorio Murino, and Nebojsa Jojic. Free energy score spaces: Using generative information in discriminative classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(7):1249–1261, 2012.

[23] Y Dan Rubinstein and Trevor Hastie. Discriminative vs Informative Learning Overview of Bayesian Classification Theory Discriminative Classification. In *Proceedings of the Third International Conference on Knowledge Discovery and Data Mining*, pages 49–53. AAAI Press, 1997.

[24] Daniel B Schwartz, Vijay K Samalam, Sara A Solla, and John S Denker. Exhaustive learning. *Neural Computation*, 2(3):374–385, September 1990.

[25] Torgyn Shaikhina and Natalia A Khovanova. Handling limited datasets with neural networks in medical applications: A small-data approach. *Artif. Intell. Med.*, 75:51–63, January 2017.

[26] Patrice Simard, Bernard Victorri, Yann LeCun, and John Denker. Tangent Prop - A formalism for specifying selected invariances in an adaptive network. In J E Moody, S J Hanson, and R P Lippmann, editors, *Advances in Neural Information Processing Systems 4*, pages 895–903. Morgan-Kaufmann, 1992.

[27] Jing Hao Xue and D. Michael Titterington. Comment on "on discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes". *Neural Processing Letters*, 28(3):169–187, 2008.

[28] Harry Zhang. The optimality of naive bayes. *Archit. Aujourdhui.*, 1(2):3, 2004.

Appendix C

Paper C

Accuracy of diagnostic classification algorithms using cognitive-, electrophysiological-, and neuroanatomical data in antipsychotic-naïve schizophrenia patients

---

Paper in press in Psychological Medicine

Ebdrup, B. H., Axelsen, M. C., Bak, N., Fagerlund, B., Oranje, B., Raghava, J. M., Nielsen, M.Ø., Rostrup, E., Hansen, L. K., Glenthøj, B. Y. (2018). Accuracy of diagnostic classification algorithms using cognitive-, electrophysiological-, and neuroanatomical data in antipsychotic-naïve schizophrenia patients. *Psychological Medicine*, 1-10.

Psychological Medicine

cambridge.org/psm

Original Article

Clinical trials identifier: NCT01154829 (registered 1 July 2010).

Cite this article: Ebdrup BH *et al* (2018). Accuracy of diagnostic classification algorithms using cognitive-, electrophysiological-, and neuroanatomical data in antipsychotic-naïve schizophrenia patients. *Psychological Medicine* 1–10. https://doi.org/10.1017/S0033291718003781

Received: 12 June 2018
Revised: 13 November 2018
Accepted: 20 November 2018

Key words:
Antipsychotic-naïve first-episode schizophrenia; cognition; diffusion tensor imaging; electrophysiology; machine learning; structural magnetic resonance imaging

Author for correspondence:
Dr Bjørn H. Ebdrup, E-mail: bebdrup@cnsr.dk

Accuracy of diagnostic classification algorithms using cognitive-, electrophysiological-, and neuroanatomical data in antipsychotic-naïve schizophrenia patients

Bjørn H. Ebdrup[1,2], Martin C. Axelsen[1,3], Nikolaj Bak[1], Birgitte Fagerlund[1,4], Bob Oranje[1,2,5], Jayachandra M. Raghava[1,6], Mette Ø. Nielsen[1,2], Egill Rostrup[1], Lars K. Hansen[3] and Birte Y. Glenthøj[1,2]

[1]Centre for Neuropsychiatric Schizophrenia Research & Centre for Clinical Intervention and Neuropsychiatric Schizophrenia Research, Mental Health Centre Glostrup, University of Copenhagen, Copenhagen, Denmark; [2]Faculty of Health and Medical Sciences, Department of Clinical Medicine, University of Copenhagen, Copenhagen, Denmark; [3]Cognitive Systems, DTU Compute, Department of Applied Mathematics and Computer Science, Technical University of Denmark, Kongens Lyngby, Denmark; [4]Department of Psychology, University of Copenhagen, Copenhagen, Denmark; [5]Department of Psychiatry, Brain Center Rudolf Magnus, University Medical Center Utrecht, Utrecht, The Netherlands and [6]Department of Clinical Physiology and Nuclear Medicine, Rigshospitalet, University of Copenhagen, Glostrup, Denmark

Abstract

Background. A wealth of clinical studies have identified objective biomarkers, which separate schizophrenia patients from healthy controls on a group level, but current diagnostic systems solely include clinical symptoms. In this study, we investigate if machine learning algorithms on multimodal data can serve as a framework for clinical translation.

Methods. Forty-six antipsychotic-naïve, first-episode schizophrenia patients and 58 controls underwent neurocognitive tests, electrophysiology, and magnetic resonance imaging (MRI). Patients underwent clinical assessments before and after 6 weeks of antipsychotic monotherapy with amisulpride. Nine configurations of different supervised machine learning algorithms were applied to first estimate the unimodal diagnostic accuracy, and next to estimate the multimodal diagnostic accuracy. Finally, we explored the predictability of symptom remission.

Results. Cognitive data significantly classified patients from controls (accuracies = 60–69%; p values = 0.0001–0.009). Accuracies of electrophysiology, structural MRI, and diffusion tensor imaging did not exceed chance level. Multimodal analyses with cognition plus any combination of one or more of the remaining three modalities did not outperform cognition alone. None of the modalities predicted symptom remission.

Conclusions. In this multivariate and multimodal study in antipsychotic-naïve patients, only cognition significantly discriminated patients from controls, and no modality appeared to predict short-term symptom remission. Overall, these findings add to the increasing call for cognition to be included in the definition of schizophrenia. To bring about the full potential of machine learning algorithms in first-episode, antipsychotic-naïve schizophrenia patients, careful *a priori* variable selection based on independent data as well as inclusion of other modalities may be required.

Introduction

A wealth of clinical studies have successfully applied various objective measures to identify biomarkers, which separate schizophrenia patients from healthy controls on a group level. Although these studies have provided profound insight into the pathophysiology of schizophrenia, these efforts have not been translated into diagnostic utility (Kapur *et al*., 2012). Thus, the diagnosis of schizophrenia according to Diagnostic and Statistical Manual of Mental Disorder (DSM) and International Classification of Diseases (ICD) classifications entirely relies on clinical symptoms. Likewise, no clinical or objective measures for course of illness or response to antipsychotic medication have been implemented into clinical practice.

Numerous studies using objective cognitive test batteries such as Cambridge Neuropsychological Test Automated Battery (CANTAB) (Robbins *et al*., 1994) have established that cognitive deficits in, e.g. attention, verbal memory, and working memory are enduring and core features of schizophrenia, which are relatively unaffected by clinical state of the psychopathological symptoms (Paulus *et al*., 2001; Gur *et al*., 2006; Kahn and Keefe, 2013).

CAMBRIDGE UNIVERSITY PRESS

Assessment of early information processing as measured with electrophysiological paradigms has also indicated impairments in schizophrenia patients, and also these disturbances are generally considered unaffected by disease stage and severity of symptoms (Koychev *et al.*, 2012; Thibaut *et al.*, 2015; Blakey *et al.*, 2018). Commonly used electrophysiological paradigms comprise P50 suppression (Adler *et al.*, 1982), pre-pulse inhibition of the startle response (PPI) (Braff and Geyer, 1990), and mismatch negativity (MMN) (Shelley *et al.*, 1991).

Finally, magnetic resonance imaging (MRI) has demonstrated that schizophrenia is associated with structural brain changes (Haijma *et al.*, 2013). Gray matter structures have commonly been assessed with a region of interest (ROI) approach, but the development of diffusion tensor imaging (DTI) techniques such as tract-based spatial statistics have enabled assessment of the cerebral white matter microstructure (Smith *et al.*, 2006). Overall, both subtle gray (Shepherd *et al.*, 2012; Gong *et al.*, 2016) and white matter (Fitzsimmons *et al.*, 2013; Canu *et al.*, 2015) deficits are present already at illness onset and before initiation of antipsychotic medication.

From a clinical perspective, the current categorical diagnostic systems contrast the multifaceted clinical phenotype of schizophrenia, and it is plausible that schizophrenia is better conceptualized using a more dimensional view (Jablensky, 2016). The research domain criteria (RDoC) were formulated to conceptualize integration of data ranging from basic biological levels to behavioral constructs across mental disorders (Insel *et al.*, 2010). Theoretically, subgroups of schizophrenia patients may share certain pathophysiological disturbances, which can serve as targets for treatment with enhanced precision (Bak *et al.*, 2017). In order to operationalize the RDoC approach, novel analysis strategies, which are sensitive to subtle signals in rich datasets, may be advantageous.

Categorical separation of groups is classically investigated with application of univariate statistical tests on unimodal data. It is increasingly appreciated that application of advanced multivariate, supervised machine learning algorithms on multimodal data may provide an improved framework for operationalizing the complex, dimensional clinical characteristics in, e.g. schizophrenia (Veronese *et al.*, 2013; Dazzan, 2014). In short, a supervised machine learning algorithm identifies 'patterns' in complex data, which are not modelled by more classical statistical methods. Next, these patterns can be used to predict the outcome (e.g. 'schizophrenia' *v.* 'healthy'; or 'remission' *v.* 'non-remission') for future, independent, individual observations with an estimated 'accuracy'. Various algorithms have been developed, each with their own advantages and disadvantages depending on, e.g. the variance and distribution of the data (Bishop, 2006; Cawley and Talbot, 2010). Previous machine learning studies have generated encouraging diagnostic accuracies >85% (e.g. Shen *et al.*, 2014; Chu *et al.*, 2016; Santos-Mayo *et al.*, 2017; Xiao *et al.*, 2017) as well as prediction of the clinical outcome (Zarogianni *et al.*, 2017). However, most previous studies have been unimodal and performed in medicated and more chronic patient samples, in which the variation in data is greater than at first illness presentation. Studies investigating multiple modalities in antipsychotic-naïve schizophrenia patients are absent.

In this proof-of-concept study, we applied nine configurations of different supervised machine learning algorithms, and we first compared the diagnostic accuracies of cognition, electrophysiology, structural MRI (sMRI), and DTI in a sample of first-episode, antipsychotic-naïve schizophrenia patients and healthy controls. Tests of group differences were supplemented with univariate analyses. Next, we investigated if combinations of modalities improved the diagnostic accuracy. Finally, we explored the predictive accuracy with regard to symptom remission after 6 weeks of antipsychotic monotherapy with amisulpride. We hypothesized that all four modalities would significantly discriminate patients from controls, and we expected higher accuracies for multimodal analyses.

Materials and methods

Trial approval

The authors assert that all procedures contributing to this work comply with the ethical standards of the Danish National Committee on Biomedical Research Ethics (H-D-2008-088) and with the Helsinki Declaration of 1975, as revised in 2008. All participants approved participation by signing informed consent. Clinical trials identifier: NCT01154829.

Participants

As part of a comprehensive multimodal study conducted between December 2008 and 2013, we recruited antipsychotic-naïve first-episode schizophrenia patients from psychiatric hospitals and outpatient mental health centers in the Capital Region of Denmark. Unimodal data on electrophysiology (Düring *et al.*, 2014, 2015), DTI (Ebdrup *et al.*, 2016), global cortical structures (Jessen *et al.*, 2018), as well as data on cognition in combination with electrophysiology (Bak *et al.*, 2017) have previously been published.

Patients were aged 18–45 years and all were lifetime naïve to any antipsychotic or methylphenidate exposure. Patients underwent a structured diagnostic interview (Schedule of Clinical Assessment in Neuropsychiatry, SCAN, version 2.1) to ensure fulfilment of ICD-10 diagnostic criteria of schizophrenia or schizoaffective psychosis (Wing *et al.*, 1990). Inclusion required a normal physical and neurological examination and no history of major head injury. Previous diagnoses of drug dependency according to ICD as well as current recreational drug use were accepted. A current diagnosis of drug dependency was an exclusion criterion. Current drug status was measured by urine test (Rapid Response, Jepsen HealthCare, Tune, Denmark). Patients treated with antidepressant medication within the last month or during the study period were excluded. Benzodiazepines and sleep medication were allowed until 12 h prior to examination days.

Duration of untreated illness (DUI) was defined as the period in which the patient reported a continuous deterioration of functioning due to disease-related symptoms (Crespo-Facorro *et al.*, 2007). Level of function was assessed with the Global Assessment of Function (GAF) and the Clinical Global Impression Scale (CGI) (Busner and Targum, 2007). Symptom severity was assessed by trained raters using the Positive and Negative Syndrome Scale (PANSS) (Kay *et al.*, 1987). After completing all baseline examinations, patients commenced amisulpride monotherapy for 6 weeks. Dosing of amisulpride was adjusted aiming to optimize clinical effect and minimize side effects. Use of anticholinergic medication was not allowed. Symptom remission after 6 weeks was assessed using the Andreasen criteria (Andreasen *et al.*, 2005).

Healthy controls matched on age, gender, and parental socioeconomic status were recruited from the community. Controls

were assessed with a SCAN interview, and former or present psychiatric illness, substance abuse, or first-degree relatives with psychiatric diagnoses, were exclusion criteria. Demographic data are presented in Table 1.

Cognition

A comprehensive neurocognitive test battery was used to assess all participants, administered by research staff trained and supervised in the standardized administration and scoring of the battery. We included variables from the following neurocognitive tasks: Danish Adult Reading Test (DART) (Nelson and O'Connell, 1978), Wechsler Adult Intelligence Scale (WAIS III) (Wechsler Adult Intelligence Scale® – Third Edition n.d.), Brief Assessment of Cognition in Schizophrenia (BACS) (Keefe *et al*., 2004), and Cambridge Neuropsychological Test Automated Battery (CANTAB) (Robbins *et al*., 1994), yielding a total of 25 cognitive variables for the current study [listed in online Supplementary Material (Table S1)].

Electrophysiology

The Copenhagen Psychophysiology Test Battery (CPTB) was used to examine all participants (Düring *et al*., 2014, 2015). Auditory stimuli were presented by a computer running 'Presentation' (Neurobehavioral Systems, Inc., Albany, NY, USA) software (soundcard: Creative soundblaster 5.1, 2008 Creative Technology Ltd, Singapore, Singapore). Stimuli were presented binaurally through stereo insert earphones (Eartone ABR, 1996–2008 Interacoustics A/S, Assens, Denmark; and C and H Distributors Inc, Milwaukee, WI, USA). To avoid cross-test influences, the CPTB is always assessed in a fixed order, including PPI, P50 suppression, MMN, and selective attention paradigms, yielding a total of 19 electrophysiological variables for the current study [listed in online Supplementary Material (Table S1)].

Neuroanatomy

MRI scans were acquired with a Philips Achieva 3.0 T whole body MRI scanner (Philips Healthcare, Best, The Netherlands) with an eight-channel SENSE Head Coil (Invivo, Orlando, Florida, USA).

Structural MRI

The three-dimensional high-resolution T1-weighted images (repetition time 10 ms, echo time 4.6 ms, flip angle 8°, voxel size 0.79 × 0.79 × 0.80 mm) were acquired and processed through FSL pipelines (Jenkinson *et al*., 2012) comprising the following steps: (1) brain extraction; (2) brain segmentation using the 'fslanat' algorithm, and resulting in gray and white matter partial volume maps for each subject; (3) non-linear warping of structural images to MNI standard space, and subsequent application of the transformation matrices to the tissue maps; (4) modulation of the warped maps using the Jacobian determinant in order to maintain local gray matter volume during the non-linear warping. Finally, regional gray matter volumes were extracted from each of the 48 anatomical regions per hemisphere derived from the Harvard–Oxford cortical atlas as specified by FSL. The total brain volume and relative ventricular volume were determined using the FSL-SIENAX program. For the brain structural analyses, we *a priori* applied the ROI approach since ROI analyses have been widely applied in the field (Haijma *et al*., 2013), and we aimed to optimize the external validity and reproducibility of the results. These procedures yielded a total of 98 sMRI variables for the current study [listed in online Supplementary Material (Table S1)].

Diffusion tensor imaging

Whole brain DTI images were acquired using single-shot spin-echo echo-planar imaging and a total of 31 different diffusion encodings [five diffusion unweighted ($b = 0$ s/mm^2) and 30 diffusion weighted ($b = 1000$ s/mm^2) non-collinear directions]. Acquired matrix size = 128 × 128 × 75; voxel dimensions = 1.88 × 1.88 mm × 2 (no slice gap); TR/TE = 7035/ 68 ms; flip angle = 90°. Images were processing using the FSL library of tools (Jenkinson *et al*., 2012). Diffusion parameter maps of fractional anisotropy (FA), mean diffusivity (MD), parallel diffusivity ($\lambda 1$), radial diffusivity ($\lambda 23$) and mode of anisotropy (MO) were derived using DTIFIT as previously described (Ebdrup *et al*., 2016). The mean values of these five diffusion parameters were extracted from 20 regions (based on the JHU white matter tractography atlas) and yielded a total of 100 DTI variables for the current study [listed in online Supplementary Material (Table S1)].

Statistical methods

Statistical Package for the Social Sciences software (version 22, SPSS Inc., USA) was used to analyze demographic and clinical data. The distribution of continuous data was tested for normality with the Shapiro–Wilk test. Data on age and years of education were not normally distributed, and group comparisons were performed non-parametrically with the Mann–Whitney U test. Group differences in gender and socioeconomic status were tested with Pearson's χ^2 test, and differences in abuse variables were tested with Fisher's exact test. Group differences in DART and estimated total IQ from WAIS III were tested using two-sample *t* tests with pooled variance estimates in MATLAB®.

Machine learning algorithms

We included participants with available data from all four modalities. We allowed subjects to have missing data points in up to 12 variables across all modalities. Twelve patients and 13 healthy controls had missing variables in the cognitive and electrophysiological data. Missing data were imputed as part of the analysis pipeline using *K*-nearest neighbor imputation with $K = 3$ (Bak and Hansen, 2016). Imputation of missing data was performed as part of the 100 random subsamples cross-validation (CV) loop, and thus the imputation procedure was only performed within the training set of a given split. We used a total of nine different configurations involving six machine learning algorithms: naïve Bayes (nB), logistic regression, support vector machine (SVM) (Cortes, 1995), decision tree (DT) (Breiman *et al*., 1984), random forest (RF) (Breiman, 2001), and auto-sklearn (AS) (Feurer *et al*., 2015). The algorithms were selected *a priori* based on their common usage and their proposed strength in relatively small datasets. To ensure comparability across all algorithms and modalities, the same pipeline and set-up were used for all analyses (Fig. 1).

Analysis pipeline

To estimate the generalization error, we used random subsampling CV (Varoquaux *et al*., 2017) with 100 stratified splits of

Table 1. Demographical and clinical data. Lifetime use of tobacco, alcohol, cannabis, stimulants, hallucinogens, and opioids were categorized according to an ordinal five-item (0 = never tried/1 = tried few times/2 = use regularly/3 = harmful use/4 = dependency)

	Schizophrenia		Healthy controls		
	N	Mean (SD)	*N*	Mean (SD)	*p*
Age, years	46	25.0 (5.6)	58	24.79(5.68)	0.79[a]
Gender (m/f)	46	28/18	58	36/22	0.901[b]
Parental SES (a/b/c)	44	8/30/6	56	16/30/10	<0.001[b]
Years of education	45	12.5 (2.6)	56	14.58(2.60)	<0.001[a]
Danish Adult Reading Test (DART)[c]	41	21.4 (9.7)	56	23.2 (6.3)	0.27[d]
Total IQ (WAIS III)[e]	41	−0.8 (1.5)	54	0.0 (1.0)	0.002[d]
Tobacco (0/1/2/3/4)	45	7/11/22/1/4	59	13/29/11/2/1	0.003[f]
Alcohol (0/1/2/3/4)	46	2/6/33/4/1	57	3/1/53/0/0	0.005[f]
Cannabis (0/1/2/3/4)	46	8/23/9/6/0	57	23/28/6/0/0	0.003[f]
Opioids (0/1/2/3/4)	46	38/8/0/0/0	56	52/4/0/0/0	0.132[f]
Stimulants (0/1/2/3/4)	46	27/14/5/0/0	56	47/9/0/0/0	0.003[f]
Hallucinogens (0/1/2/3/4)	45	38/7/0/0/0	57	53/3/0/0/0	0.105[f]
Other drugs (0/1/2/3/4)	43	40/3/0/0/0	56	54/2/0/0/0	0.65[f]
Benzodiazepines (0/1/2/3/4)	42	31/11/0/6/0	55	55/0/0/0/0	<0.001[f]
DUI, weeks	45	65.8 (70.5)	–	–	–
CGI, severity	44	4.2 (0.7)	–	–	–
GAF, symptom	44	40.9 (9.9)	–	–	–
GAF, function	43	42.6 (11.1)	–	–	–
PANSS, positive	46	20.1 (4.2)	–	–	–
PANSS, negative	46	21.3 (7.9)	–	–	–
PANSS, general	46	42.2 (9.4)	–	–	–
PANSS, total	46	83.5 (17.2)	–	–	–
Amisulpride, mg/day	32	248.4 (140.6)	–	–	–
Remission (yes/no)[g]	34	11/23	–	–	–

SES, parental socioeconomic status; DUI, duration of untreated illness; CGI, Clinical Global Impression Scale; GAF, Global Assessment of Functioning; PANSS, Positive And Negative Syndrome Scale.
[a]Mann–Whitney *U* test.
[b]χ^2.
[c]Danish Adult Reading Test (DART) (Nelson and O'Connell, 1978).
[d]Two-sample *t* test with pooled variance estimates.
[e]A combined score based on four subtests from WAIS III: Wechsler Adult Intelligence Scale (Wechsler Adult Intelligence Scale® – Third Edition n.d.), presented as *z*-scores standardized from the mean and standard deviation of the healthy control sample.
[f]Fisher's exact test.
[g]Symptom remission after 6 weeks according to Andreasen criteria (Andreasen *et al.*, 2005).

patients and controls (Fig. 1). This approach ensured that all configurations of algorithms were trained on the same data, and the ratio between the two classes was similar for all splits. Therefore, the performance of algorithms was evaluated on the same test data. For each split, one-third of the data was used for testing and two-thirds were used for training. All data imputation, feature selection, model training, and optimization were based exclusively on the training set of a given split. Logistic regression was used in two configurations: with L1 regularization (LR_r) and without regularization (LR). SVM was used in three configurations: one with a linear kernel (SVM_l), one with a radial basis function kernel using heuristic parameters (SVM_h), and one with optimized parameters (SVM_o). An inner loop fivefold CV was used to optimize model parameters (LR_r, SVM_o) or perform backwards elimination feature selection (LR, SVM_l, SVM_h, DT). Algorithms RF and AS have inherent parameter optimization, and therefore these configurations required no inner loop CV. See online Supplementary Material 'Machine learning algorithms' for details.

Strategy for analyses

To acquire unimodal estimates for the ability to separate patients from healthy controls (i.e. the 'diagnostic accuracy'), data from each of the four modalities (cognition, electrophysiology, sMRI, and DTI) were analyzed using each of the nine configurations of machine learning algorithms yielding nine estimates per modality (Fig. 2). In order to compare the contribution of

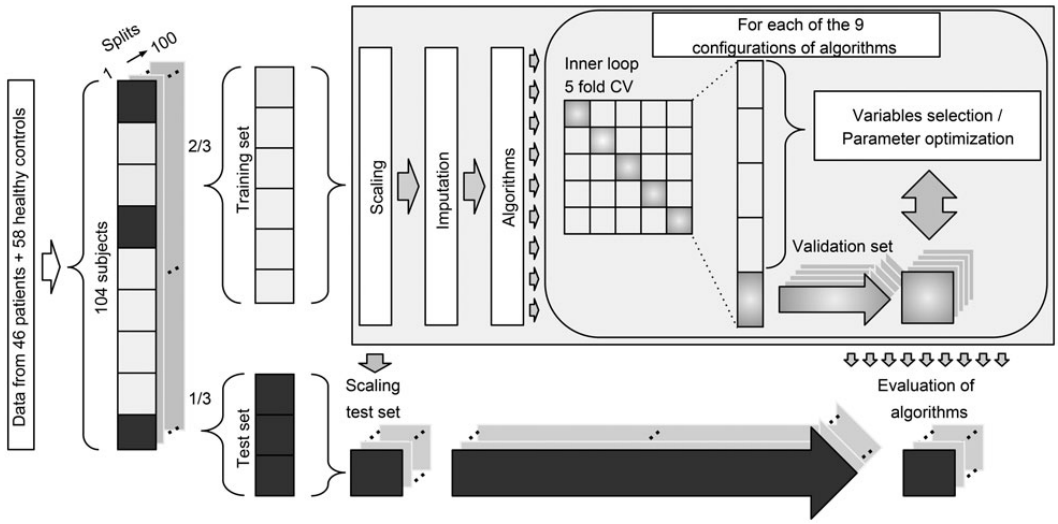

Fig. 1. Diagram of the multivariate analysis pipeline. Forty-six patients and 58 healthy controls were included in the baseline analyses. 'Data' refer to input variables from cognition, electrophysiology, structural magnetic resonance imaging, and diffusion tensor imaging. For each of the 100 splits, 2/3 of subjects were used for training and 1/3 of subjects were used for testing. Subjects with missing data were not used in test sets. Training data were scaled (zero mean, unit variance), and the test sets were scaled using these parameters. Missing data were imputed using *K*-nearest neighbor imputation with $K = 3$ (Bak and Hansen, 2016), and only subjects with complete data were included in the test sets. Finally, nine different configurations of machine learning algorithms were applied to predict diagnosis. CV = cross-validation. See text for details.

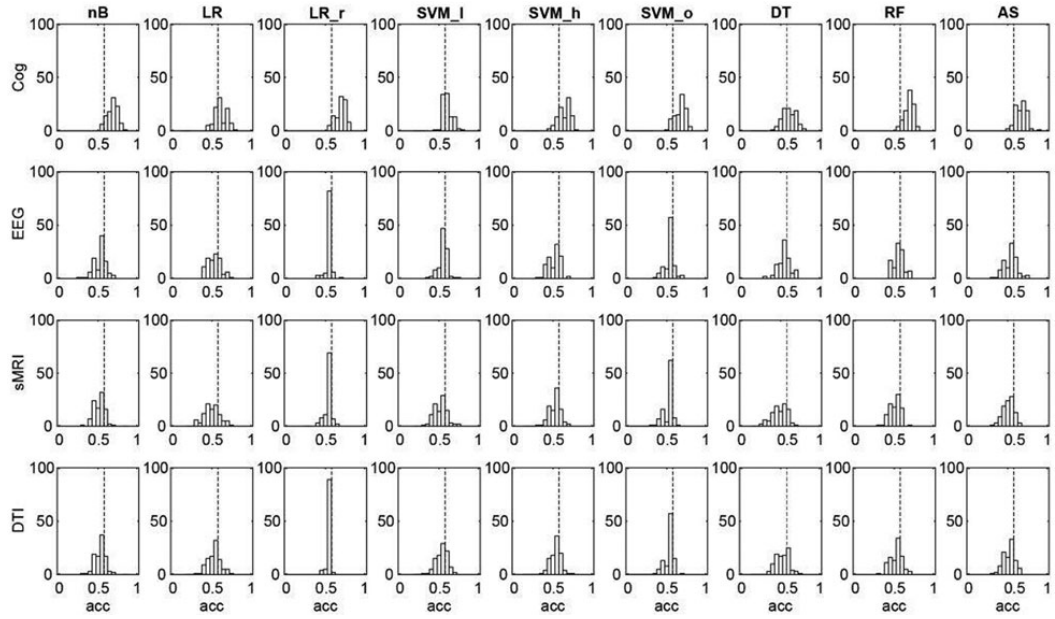

Fig. 2. Unimodal diagnostic accuracies for cognition (Cog), electrophysiology (EEG), structural magnetic resonance imaging (sMRI), and diffusion tensor imaging (DTI) for each of the nine different configurations of machine learning algorithms. *X*-axes show the accuracies (acc), and *y*-axes show the sum of correct classifications for each of the 100 random subsamples (see Fig. 1). Dotted vertical black line indicates chance accuracy (56%). With cognitive data, all nine configurations of algorithms significantly classified 'patient *v.* control' (p values = 0.001–0.009). No algorithms using EEG, sMRI, and DTI-data resulted in accuracies exceeding chance. The nine different configuration of machine learning algorithms: nB, naïve Bayes; LR, logistic regression without regularization; LR_r, logistic regression with regularization; SVM_l, support vector machine with linear kernel; SVM_h, SVM with heuristic parameters; SVM_o, SVM optimized through cross-validation; DT, decision tree; RF, random forest; AS, auto-sklearn. See text for details.

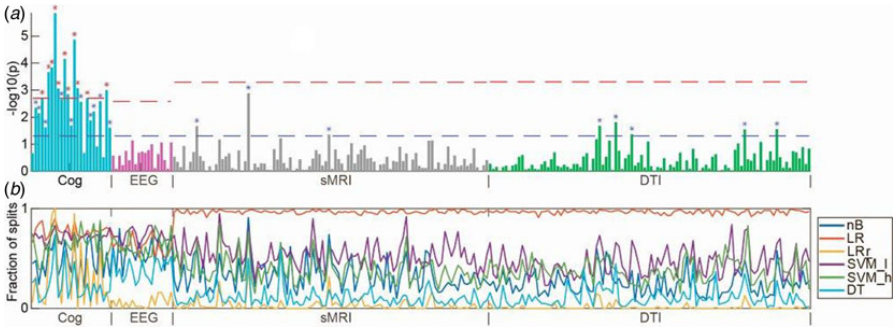

Fig. 3. (*a*) Manhattan plot with univariate *t* tests of all variables along the *x*-axis [cognition (Cog), electrophysiology (EEG), structural magnetic resonance imaging (sMRI), and diffusion tensor imaging (DTI)] and log-transformed *p* values along the *y*-axis. Lower dashed horizontal line indicates significance level of $p = 0.05$. Upper dashed lines indicate the Bonferroni-corrected *p* value for each modality. (*b*) In colored horizontal lines, the fraction of data splits (see Fig. 1), where individual variables were included in the final machine learning model, which determined the diagnostic accuracy (presented in Fig. 2). Specification of variables is provided in online Supplementary Material. Only configurations of the six machine learning algorithms, which included feature selection, are shown. nB, naïve Bayes; LR, logistic regression without regularization; LR_r, logistic regression with regularization; SVM_l, support vector machine with linear kernel; DT, decision tree; RF, random forest.

individual variables to these unimodal multivariate estimates, we performed univariate *t* tests between patients and healthy controls (Fig. 3). In order to estimate the multimodal diagnostic accuracy, any modality, which significantly discriminated between patients and healthy controls, was analyzed with all seven combinations in an early integration of the remaining modalities, where variables are concatenated to form larger combined modalities. Finally, we explored if any modality predicted PANSS symptom remission according to the Andreasen criteria (Andreasen *et al.*, 2005). Analyses of symptom remission were performed for patients only, and for these analyses, a fifth 'clinical modality' was constructed. The clinical modality comprised basic demographic and clinical features, which may influence on illness prognosis: age, gender substance use, DUI, GAF (symptoms and function), and PANSS subscores (positive, negative, and general symptoms). To estimate prediction of symptom remission after 6 weeks of amisulpride treatment, data from each of the five modalities were analyzed using all nine configurations of algorithms via the same analysis pipeline as described above (Fig. 1).

Results

Demographics

Forty-six patients and 58 healthy controls were included in the current analyses. Groups were well matched on age, gender, but parental socioeconomic status was lower in patients compared with controls. Compared with controls, the patients had significantly fewer years of education, and significantly higher use of tobacco and recreational drugs, except for use of hallucinogens. Patients were treated with amisulpride in an average dose of 248.4 mg/day for 6 weeks. After 6 weeks of amisulpride treatment, 11 out of 34 (32%) patients fulfilled remission criteria (Andreasen *et al.*, 2005) (Table 1).

Unimodal diagnostic accuracy

Since the two groups differed in size (46 patients and 58 healthy controls), the 'chance accuracy' was 56% $[(58/(46 + 58) \times 100)]$. The diagnostic accuracy of cognition ranged between 60% and 69% for all nine configurations of algorithms. A permutation test using 1000 permutations showed that all configurations using cognitive data significantly differentiated between patients and controls (p values ranging from 0.001 to 0.009) (see online Supplementary Material, Table S1). The diagnostic accuracy for electrophysiology, sMRI, and DTI ranged between 49% and 56% and did not exceed chance accuracy (Fig. 2).

The planned *t* tests showed that 11/25 of the cognitive variables survived Bonferroni correction ($0.05/25 = 0.002$) (Fig. 3). The variables covered domains of IQ, working memory, motor function, verbal fluency, processing speed, executive functions, spatial working memory, and sustained attention (see online Supplementary Material, Table S2 for specification of variables). None of 19 electrophysiological, 3/98 sMRI, and 5/100 DTI variables significantly differed between patients and controls at $p < 0.05$; however, none survived after Bonferroni correction (Fig. 3 and online Supplementary Material, Table S2).

Multimodal diagnostic accuracy

None of the multimodal analyses with cognition plus any combination of one or more of the remaining modalities (electrophysiology, sMRI, and DTI) revealed significantly higher accuracies than cognition alone (accuracies ranging between 51% and 68%) (see online Supplementary Material, Table S1).

Prognostic ability

Using symptomatic remission ($N = 11$) *v.* non-remission ($N = 23$) as a dichotomous outcome measure equals a 'chance accuracy' of 68% $[(23/(11 + 23) \times 100)]$. None of the modalities predicted symptom remission after 6 weeks above chance level: cognition, electrophysiology, sMRI, and DTI predicted symptom remission at accuracies ranging between 48% and 67%. The fifth 'clinical variable' predicted symptom remission with accuracies ranging between 51% and 67% (see online Supplementary Material, Table S3).

Discussion

To our knowledge, this is the first study to investigate the diagnostic accuracy of machine learning algorithms using multimodal

data in antipsychotic-naïve, first-episode schizophrenia patients. Contrary to our expectations, we found that only cognitive data, but no other modality, significantly discriminated patients from healthy controls. Moreover, we did not find enhanced accuracies by combining cognition with other modalities, and finally, none of the modalities predicted symptom remission.

Based on cognitive data, all nine configurations of machine learning algorithms could separate patients from healthy controls with a statistically significant accuracy. Supervised machine learning algorithms model the interdependent pattern of variables, which best separate the data with respect to the outcome (e.g. 'schizophrenia' or 'healthy'). Our *t* tests indicated that patients differed from controls on a broad spectrum of cognitive domains, and the feature selection lines shown in Fig. 3*b* indicate that variables with lower *p* values were included more frequently in the machine learning models. Hence, at initial diagnosis of schizophrenia, cognitive deficits appear markedly more pronounced than electrophysiological and neuroanatomical aberrations. Interestingly, two previous multimodal studies in medicated patients also indicated that cognitive parameters yielded higher classification accuracies than sMRI (Karageorgiou *et al*., 2011), and genotype, DTI, and fMRI (Pettersson-Yeo *et al*., 2013). Cognitive deficits are not a part of the diagnostic criteria for schizophrenia, although this has been discussed in the field before the implementation of DSM-5 (Kahn and Keefe, 2013). Our findings support resuming these discussions and examining the evidence for including objective cognitive assessment into future diagnostic systems.

The accuracies regarding neuroanatomical and electrophysiological markers reported in this study are remarkably lower than the accuracies reported in several previous studies. A recent meta-analysis of 20 sMRI studies concluded that application of multivariate algorithms could discriminate schizophrenia patients from healthy controls with a sensitivity of 76% and a specificity of 79% (Kambeitz *et al*., 2015). Higher age and more psychotic symptoms, which in turn may be associated with illness duration and illness severity, more antipsychotic exposure, and more substance abuse, were identified as significant moderators. Moreover, resting-state fMRI data were superior to sMRI in discriminating schizophrenia patients from controls. In the current study, patients were all antipsychotic-naïve, relatively young (mean age of 25.0 years), and displayed moderate psychotic symptoms (PANSS-positive symptoms of 20.1) (Table 1). Furthermore, resting-state fMRI was not included. A previous study using electrophysiological data from 16 schizophrenia patients and 31 healthy controls resulted in a correct classification rate of around 93%. Notably, different EEG measures were used than in the current study, and a mean age of 36 years suggests that the patients were chronically ill and medicated (Santos-Mayo *et al*., 2017). Collectively, the limited clinical confounders in the current study may have contributed to the low diagnostic accuracies of sMRI and DTI, and electrophysiology.

Moreover, methodological differences may contribute to explain the current findings. To optimize the external validity, we applied a rigorous approach in our analysis pipeline. Specifically, we used all available variables, i.e. no feature selection was done prior to entering data into the analysis pipeline. Generally, the studies, which have reported very high accuracies, have first applied a statistical test to pre-select variables, which discriminate between groups on the outcome measure for the specific dataset (e.g. Chu *et al*., 2016; Santos-Mayo *et al*., 2017). A recent SVM study using sMRI cortical thickness and surface data from 163 first-episode, antipsychotic-naïve patients (mean age 23.5 years) and matched controls (mean age 23.6 years) revealed a diagnostic accuracy of 81.8% and 85.0%, respectively, for thickness and surface. In that study, the SVM input comprised variables, which separated patients from controls on a *t* test adjusted for multiple comparisons (Xiao *et al*., 2017). Conversely, a recent machine learning study on voxel-based MRI data from 229 schizophrenia patients and 220 healthy controls from three independent datasets used no prior feature selection and reported low accuracies ranging between 55% and 73.5% (Winterburn *et al*., 2017). Thus, pre-analysis feature selection may provide higher accuracies at the expense of generalizability of the results and should therefore be discouraged in studies aiming at clinical translation.

Contrary to our expectations, we did not find added diagnostic accuracy when combining cognition with other modalities. Moreover, neither cognition nor our constructed 'clinical variable' predicted symptom remission after 6 weeks according to criteria which were validated after 6 months of treatment (Andreasen *et al*., 2005). Since the between-subject variability in our data is large, but the group differences between antipsychotic-naïve patients and healthy controls regarding electrophysiology and neuroanatomy are subtle, our results encourage application of multimodal, multivariate analyses in order to disentangle neurobiological distinct subgroups within cohorts of schizophrenia patients. Specifically, multimodal, multivariate analyses may identify clinically meaningful subgroups of schizophrenia patient, e.g. with regard to clinical trajectories (Bak *et al*., 2017). Finally, and in line with the RDoC initiative, it is conceivable that indices of clinical trajectories may expand beyond psychopathology also to encompass more objective, biologically valid assessments.

Some strengths and limitations should be considered. At inclusion, the patients were antipsychotic-naïve and as intervention we used a relatively selective dopamine D_2 receptor antagonist. Therefore, our diagnostic accuracies reflect minimally confounded estimates of neurobiological disturbances at the earliest stage of schizophrenia. First-episode, antipsychotic-naïve patients are challenging to recruit, and since we required close to complete datasets from all participants on four modalities, the number of included patients may have been too small for optimal modeling of electrophysiology, sMRI, and DTI data. The four modalities used in this study were *a priori* selected because our own eletrophysiological (Düring *et al*., 2014, 2015) and DTI data (Ebdrup *et al*., 2016) as well as abundant independent data have rather consistently shown group differences between schizophrenia patients and controls. Moreover, data on these four modalities can be obtained by means of relatively standardized procedures, which enhances the generalizability our study. As we have also previously published group differences on this cohort in reward processing (Nielsen *et al*., 2012*a*, 2012*b*), resting-state activity (Anhøj *et al*., 2018), and striatal dopamine D_2 receptor-binding potentials (Wulff *et al*., 2015), inclusion of functional MRI or neurochemical data may have given more positive results. In the current study, we aimed at balancing measures with high clinical generalizability on the largest possible dataset. Because of the absence of standardized pipelines for more dynamic and task-dependent measures, and because inclusion of additional modalities would have reduced the number of participant with full datasets, we *a priori* decided not to include fMRI and neurochemical data in the current analyses. Nevertheless, across all four modalities, our nine different configurations of machine learning algorithms appeared to detect similar signals as the

conventional *t* tests (Fig. 3*b*). This overlap in signal provides indirect validation of the applied methods and implies that multivariate algorithms are not a 'black box' (Castelvecchi, 2016). As recommended in a recent meta-analysis of machine learning classifications studies, we corrected for age and demographical group differences (Neuhaus and Popescu, 2018). Nevertheless, our modest sample size requires replication in an independent sample, which was currently not available. Regarding prediction of outcome, we only evaluated symptom remission with respect to criteria, which were validated after 6 months of treatment (Andreasen *et al.*, 2005). Because our analyses of symptom remission were based on only 34 patients (11 patients were in remission), these results should also be interpreted cautiously since we cannot exclude a Type 2 error.

The inclusion of all available data resulted in an unintended group difference in parental socioeconomic status (Table 1). There were no group differences in premorbid IQ (i.e. DART), but significant group differences on estimated total IQ, with effect sizes similar to previous findings in first-episode samples (Mesholam-Gately *et al.*, 2009), but still, these sociodemographic differences cannot explain the marked group differences in cognitive performance we see between groups. We allowed benzodiazepines on an 'as needed' basis until 12 h prior to examination days to reduce anxiety and secure sleep. Therefore, we cannot exclude an effect of benzodiazepines on our results; however, since sleep restriction also negatively affects cognition (Lowe *et al.*, 2017), we judge the potential bias of benzodiazepines minimal. Our comprehensive approach where we included all available variables may have compromised the signal-to-noise ratio. *A priori* selection of predefined candidate variables, i.e. to make use of 'domain knowledge', could potentially have enhanced our signal-to-noise ratio, and in turn our accuracies, without compromising the external validity. Moreover, for neuroanatomical analyses, we included regions of interest. Although a voxel-based approach may be more sensitive to global brain structural aberrations, this was not the case in the recent large machine learning study on voxel-based MRI data mentioned above (Winterburn *et al.*, 2017).

Visual inspection of the *t* tests presented in Fig. 3*a* show that the magnitude of cognitive group differences is marked and extensive (22/25 variables had p values <0.05), whereas only few variables from electrophysiology, sMRI, and DTI had p values <0.05. A more liberal correction for multiple comparisons than the applied Bonferroni correction, e.g. the false discovery rate *ad modum* Benjamini–Hochberg (Benjamini and Hochberg, 1995) would not have changed our overall conclusion that cognitive deficits, compared with electrophysiological and regional brain measures, are core features of schizophrenia at first clinical presentation (Kahn and Keefe, 2013). Since we only investigated one diagnostic category (i.e. schizophrenia), we cannot infer to what extent the discriminative diagnostic patterns of cognitive disturbances are specific to schizophrenia *per se* (Bora and Pantelis, 2016).

In conclusion, this multivariate and multimodal proof-of-concept study on antipsychotic-naïve patients showed that cognition, but not electrophysiological and neuroanatomical data, significantly discriminated schizophrenia patients from healthy controls. Overall, these findings add to the increasing call for cognition to be included in the definition of schizophrenia. To bring about the full potential of machine learning algorithms in first-episode, antipsychotic-naïve schizophrenia patients, careful a *priori* variable selection based on independent data as well as inclusion of other modalities may be required. Machine learning studies aiming at identification of clinically meaningful subgroups of schizophrenia patients are encouraged.

Supplementary material. The supplementary material for this article can be found at https://doi.org/10.1017/S0033291718003781.

Acknowledgements. None.

Financial support. The study was supported by unrestricted grant R25-A2701 from the Lundbeck Foundation to the Centre for Clinical Intervention and Neuropsychiatric Schizophrenia Research (CINS).

Conflict of interest. Dr BE has received lecture fees and/or is part of Advisory Boards of Bristol-Myers Squibb, Eli Lilly and Company, Janssen-Cilag, Otsuka Pharma Scandinavia, Lundbeck Pharma A/S, and Takeda Pharmaceutical Company. Dr NB became a full-time employee at Lundbeck Pharma A/S, Denmark after completion of this study. All other authors report no conflicts of interest.

Author ORCIDs. Bjørn H. Ebdrup 0000-0002-2590-5055

References

Adler LE, Pachtman E, Franks RD, Pecevich M, Waldo MC and Freedman R (1982) Neurophysiological evidence for a defect in neuronal mechanisms involved in sensory gating in schizophrenia. *Biological Psychiatry* **17**, 639–654.

Andreasen NC, Carpenter WT, Kane JM, Lasser RA, Marder SR and Weinberger DR (2005) Remission in schizophrenia: proposed criteria and rationale for consensus. *American Journal of Psychiatry* **162**, 441–449.

Anhøj S, Ødegaard Nielsen M, Jensen MH, Ford K, Fagerlund B, Williamson P, Glenthøj B and Rostrup E (2018) Alterations of intrinsic connectivity networks in antipsychotic-naïve first-episode schizophrenia. *Schizophrenia Bulletin* **44**, 1332–1340.

Bak N and Hansen LK (2016) Data driven estimation of imputation error – a strategy for imputation with a reject option. Ed. Z Zhang. *PLoS ONE* **11**, e0164464.

Bak N, Ebdrup BHH, Oranje B, Fagerlund B, Jensen MHH, Düring SWW, Nielsen MØØ, Glenthøj BYY and Hansen LKK (2017) Two subgroups of antipsychotic-naive, first-episode schizophrenia patients identified with a Gaussian mixture model on cognition and electrophysiology. *Translational Psychiatry* **7**, e1087.

Benjamini Y and Hochberg Y (1995) Controlling the false discovery rate: a Practical and powerful approach to Multiple testing. Wiley Royal Statistical Society. *Journal of the Royal Statistical Society. Series B* (*Methodological*) **57**, 289–300.

Bishop CM (2006) *Pattern Recognition and Machine Learning*. New York, NY: Springer.

Blakey R, Ranlund S, Zartaloudi E, Cahn W, Calafato S, Colizzi M, Crespo-Facorro B, Daniel C, Díez-Revuelta Á, Di Forti M, Iyegbe C, Jablensky A, Jones R, Hall M-H, Kahn R, Kalaydjieva L, Kravariti E, Lin K, McDonald C, McIntosh AM, Picchioni M, Powell J, Presman A, Rujescu D, Schulze K, Shaikh M, Thygesen JH, Toulopoulou T, Van Haren N, Van Os J, Walshe M, Murray RM, Bramon E and Bramon E (2018) Associations between psychosis endophenotypes across brain functional, structural, and cognitive domains. *Psychological Medicine* **48**, 1325–1340.

Bora E and Pantelis C (2016) Social cognition in schizophrenia in comparison to bipolar disorder: a meta-analysis. *Schizophrenia Research* **175**, 72–78.

Braff DL and Geyer MA (1990) Sensorimotor gating and schizophrenia. Human and animal model studies. *Archives of general Psychiatry* **47**, 181–188.

Breiman L (2001) Random forests. Kluwer Academic Publishers *Machine Learning* **45**, 5–32.

Breiman L, Friedman J, Olshen R and Stone C (1984) *Classification and Regression Trees*. New York: Chapman and Hall, Wadsworth.

Busner J and Targum SD (2007) The clinical global impressions scale: applying a research tool in clinical practice. Matrix Medical Communications Psychiatry (Edgmont (Pa. : Township)) **4**, 28–37.

Canu E, Agosta F and Filippi M (2015) A selective review of structural connectivity abnormalities of schizophrenic patients at different stages of the disease. *Schizophrenia Research* **161**, 19–28.

Castelvecchi D (2016) Can we open the black box of AI? *Nature* **538**, 20–23.

Cawley G and Talbot N (2010) On over-fitting in model selection and subsequent selection bias in performance evaluation. *Journal of Machine Learning Research* **11**, 2079–2107.

Chu W-L, Huang M-W, Jian B-L, Hsu C-Y and Cheng K-S (2016) A correlative classification study of schizophrenic patients with results of clinical evaluation and structural magnetic resonance images. *Behavioural Neurology* **2016**, 1–11.

Cortes C (1995). *Support-Vector Networks*. vol 20.

Crespo-Facorro B, Roiz-Santiáñez R, Pelayo-Terán JM, González-Blanch C, Pérez-Iglesias R, Gutiérrez A, de Lucas EM, Tordesillas D and Vázquez-Barquero JL (2007) Caudate nucleus volume and its clinical and cognitive correlations in first episode schizophrenia. *Schizophrenia Research* **91**, 87–96.

Dazzan P (2014) Neuroimaging biomarkers to predict treatment response in schizophrenia: the end of 30 years of solitude? *Dialogues in Clinical Neuroscience* **16**, 491–503.

Düring S, Glenthøj BY, Andersen GS and Oranje B (2014) Effects of dopamine D2/D3 blockade on human sensory and sensorimotor gating in initially antipsychotic-naive, first-episode schizophrenia patients. *Neuropsychopharmacology* **39**, 3000–3008.

Düring S, Glenthøj BY and Oranje B (2015) Effects of blocking D2/D3 receptors on mismatch negativity and P3a amplitude of initially antipsychotic naïve, first episode schizophrenia patients. *The International Journal of Neuropsychopharmacology* **19**, pyv109.

Ebdrup BH, Raghava JM, Nielsen MØ, Rostrup E and Glenthøj B (2016) Frontal fasciculi and psychotic symptoms in antipsychotic-naive patients with schizophrenia before and after 6 weeks of selective dopamine D2/3 receptor blockade. *Journal of Psychiatry & Neuroscience : JPN* **41**, 133–141.

Feurer M, Klein A, Eggensperger K, Springenberg J, Blum M and Hutter F (2015) *Efficient and Robust Automated Machine Learning* 2962–2970.

Fitzsimmons J, Kubicki M and Shenton ME (2013) Review of functional and anatomical brain connectivity findings in schizophrenia. *Current Opinion in Psychiatry* **26**, 172–187.

Gong Q, Lui S and Sweeney JA (2016) A selective review of cerebral abnormalities in patients with first-episode schizophrenia before and after treatment. *The American Journal of Psychiatry* **173**, 232–243.

Gur RE, Calkins ME, Gur RC, Horan WP, Nuechterlein KH, Seidman LJ and Stone WS (2006) The consortium on the genetics of schizophrenia: neurocognitive endophenotypes. *Schizophrenia Bulletin* **33**, 49–68.

Haijma SV, Van Haren N, Cahn W, Koolschijn PCMP, Hulshoff Pol HE and Kahn RS (2013) Brain volumes in schizophrenia: a meta-analysis in over 18 000 subjects. *Schizophrenia Bulletin* **39**, 1129–1138.

Insel T, Cuthbert B, Garvey M, Heinssen R, Pine DS, Quinn K, Sanislow C and Wang P (2010) Research domain criteria (RDoC): toward a new classification framework for research on mental disorders. *American Journal of Psychiatry* **167**, 748–751.

Jablensky A (2016) Psychiatric classifications: validity and utility. *World Psychiatry* **15**, 26–31.

Jenkinson M, Beckmann CF, Behrens TEJ, Woolrich MW and Smith SM (2012) FSL. *NeuroImage* **62**, 782–790.

Jessen K, Rostrup E, Mandl RCW, Nielsen MØ, Bak N, Fagerlund B, Glenthøj BY and Ebdrup BH (2018) Cortical structures and their clinical correlates in antipsychotic-naïve schizophrenia patients before and after 6 weeks of dopamine $D_{2/3}$ receptor antagonist treatment. *Psychological Medicine* **8**, 1-10.

Kahn RS and Keefe RSE (2013). Schizophrenia is a cognitive illness: time for a change in focus. *JAMA Psychiatry* **70**, 1107–1112.

Kambeitz J, Kambeitz-Ilankovic L, Leucht S, Wood S, Davatzikos C, Malchow B, Falkai P and Koutsouleris N (2015). Detecting neuroimaging biomarkers for schizophrenia: a meta-analysis of multivariate pattern recognition studies. *Neuropsychopharmacology* **40**, 1742–1751.

Kapur S, Phillips AG and Insel TR (2012). Why has it taken so long for biological psychiatry to develop clinical tests and what to do about it? *Molecular Psychiatry* **17**, 1174–1179.

Karageorgiou E, Schulz SC, Gollub RL, Andreasen NC, Ho B-C, Lauriello J, Calhoun VD, Bockholt HJ, Sponheim SR and Georgopoulos AP (2011) Neuropsychological testing and structural magnetic resonance imaging as diagnostic biomarkers early in the course of schizophrenia and related psychoses. *Neuroinformatics* **9**, 321–333.

Kay SR, Fiszbein A and Opler LA (1987) The positive and negative syndrome scale (PANSS) for schizophrenia. *Schizophrenia Bulletin* **13**, 261–276.

Keefe RSE, Goldberg TE, Harvey PD, Gold JM, Poe MP and Coughenour L (2004) The Brief Assessment of Cognition in Schizophrenia: reliability, sensitivity, and comparison with a standard neurocognitive battery. *Schizophrenia Research* **68**, 283–297.

Koychev I, El-Deredy W, Mukherjee T, Haenschel C and Deakin JFW (2012) Core dysfunction in schizophrenia: electrophysiology trait biomarkers. *Acta Psychiatrica Scandinavica* **126**, 59–71.

Lowe CJ, Safati A and Hall PA (2017) The neurocognitive consequences of sleep restriction: a meta-analytic review. *Neuroscience & Biobehavioral Reviews* **80**, 586–604.

Mesholam-Gately RI, Giuliano AJ, Goff KP, Faraone SV and Seidman LJ (2009). Neurocognition in first-episode schizophrenia: a meta-analytic review. *Neuropsychology* **23**, 315–336.

Nelson HE and O'Connell A (1978) Dementia: the estimation of premorbid intelligence levels using the New Adult Reading Test. *Cortex* **14**, 234–244.

Neuhaus AH and Popescu FC (2018) Impact of sample size and matching on single-subject classification of schizophrenia: a meta-analysis. *Schizophrenia Research* **192**, 479–480.

Nielsen MO, Rostrup E, Wulff S, Bak N, Broberg BV, Lublin H, Kapur S and Glenthoj B (2012*a*) Improvement of brain reward abnormalities by antipsychotic monotherapy in schizophrenia. *Archives of General Psychiatry* **69**, 1195–1204.

Nielsen MØ, Rostrup E, Wulff S, Bak N, Lublin H, Kapur S and Glenthøj B (2012*b*) Alterations of the brain reward system in antipsychotic naïve schizophrenia patients. *Biological Psychiatry* **71**, 898–905.

Paulus MP, Rapaport MH and Braff DL (2001) Trait contributions of complex dysregulated behavioral organization in schizophrenic patients. *Biological Psychiatry* **49**, 71–77.

Pettersson-Yeo W, Benetti S, Marquand AF, Dell'Acqua F, Williams SCR, Allen P, Prata D, McGuire P and Mechelli A (2013) Using genetic, cognitive and multi-modal neuroimaging data to identify ultra-high-risk and first-episode psychosis at the individual level. *Psychological Medicine* **43**, 2547–2562.

Robbins TW, James M, Owen AM, Sahakian BJ, McInnes L and Rabbitt P (1994) Cambridge Neuropsychological Test Automated Battery (CANTAB): a factor analytic study of a large sample of normal elderly volunteers. *Dementia (Basel, Switzerland)* **5**, 266–281.

Santos-Mayo L, San-Jose-Revuelta LM and Arribas JI (2017) A computer-aided diagnosis system with EEG based on the P3b wave during an auditory odd-ball task in schizophrenia. *IEEE Transactions on Biomedical Engineering* **64**, 395–407.

Shelley AM, Ward PB, Catts SV, Michie PT, Andrews S and McConaghy N (1991) Mismatch negativity: an index of a preattentive processing deficit in schizophrenia. *Biological Psychiatry* **30**, 1059–1062.

Shen C, Popescu FC, Hahn E, Ta TTM, Dettling M and Neuhaus AH (2014) Neurocognitive pattern analysis reveals classificatory hierarchy of attention deficits in schizophrenia. Oxford University Press *Schizophrenia Bulletin* **40**, 878–885.

Shepherd AM, Laurens KR, Matheson SL, Carr VJ and Green MJ (2012) Systematic meta-review and quality assessment of the structural brain alterations in schizophrenia. *Neuroscience and Biobehavioral Reviews* **36**, 1342–1356.

Smith SM, Jenkinson M, Johansen-Berg H, Rueckert D, Nichols TE, Mackay CE, Watkins KE, Ciccarelli O, Cader MZ, Matthews PM and Behrens TEJ (2006) Tract-based spatial statistics: voxelwise analysis of multi-subject diffusion data. *NeuroImage* **31**, 1487–1505.

Thibaut F, Boutros NN, Jarema M, Oranje B, Hasan A, Daskalakis ZJ, Wichniak A, Schmitt A, Riederer P and Falkai P, WFSBP Task Force on Biological Markers (2015) Consensus paper of the WFSBP Task Force on Biological Markers: criteria for biomarkers and endophenotypes of schizophrenia part I: neurophysiology. *The World Journal of Biological Psychiatry* **16**, 280–290.

Varoquaux G, Raamana PR, Engemann DA, Hoyos-Idrobo A, Schwartz Y and Thirion B (2017) Assessing and tuning brain decoders: cross-validation, caveats, and guidelines. *NeuroImage* **145**, 166–179.

Veronese E, Castellani U, Peruzzo D, Bellani M and Brambilla P (2013) Machine learning approaches: from theory to application in schizophrenia. *Computational and Mathematical Methods in Medicine* **2013**, 867924.

Wechsler Adult Intelligence Scale® – Third Edition (n.d.).

Wing JK, Babor T, Brugha T, Burke J, Cooper JE, Giel R, Jablenski A, Regier D and Sartorius N (1990) SCAN. Schedules for Clinical Assessment in Neuropsychiatry. *Archives of General Psychiatry* **47**, 589–593.

Winterburn JL, Voineskos AN, Devenyi GA, Plitman E, de la Fuente-Sandoval C, Bhagwat N, Graff-Guerrero A, Knight J and Chakravarty MM (2017) Can we accurately classify schizophrenia patients from healthy controls using magnetic resonance imaging and machine learning? A multi-method and multi-dataset study. *Schizophrenia Research*. doi: 10.1016/j.schres.2017.11.038.

Wulff S, Pinborg LH, Svarer C, Jensen LT, Nielsen MØ, Allerup P, Bak N, Rasmussen H, Frandsen E, Rostrup E and Glenthøj BY (2015) Striatal D (2/3) binding potential values in drug-naïve first-episode schizophrenia patients correlate with treatment outcome. *Schizophrenia Bulletin* **41**, 1143–1152.

Xiao Y, Yan Z, Zhao Y, Tao B, Sun H, Li F, Yao L, Zhang W, Chandan S, Liu J, Gong Q, Sweeney JA and Lui S (2017) Support vector machine-based classification of first episode drug-naïve schizophrenia patients and healthy controls using structural MRI. *Schizophrenia Research*. doi: 10.1016/j.schres.2017.11.037.

Zarogianni E, Storkey AJ, Johnstone EC, Owens DGC and Lawrie SM (2017) Improved individualized prediction of schizophrenia in subjects at familial high risk, based on neuroanatomical data, schizotypal and neurocognitive features. *Schizophrenia Research* **181**, 6–12.

Bibliography

[Abrahamsen and Hansen, 2011] Abrahamsen, T. J. and Hansen, L. K. (2011). A cure for variance inflation in high dimensional kernel principal component analysis. *J. Mach. Learn. Res.*, 12:2027–2044.

[Aggernaes et al., 2010] Aggernaes, B., Glenthoj, B. Y., Ebdrup, B. H., Rasmussen, H., Lublin, H., and Oranje, B. (2010). Sensorimotor gating and habituation in antipsychotic-naive, first-episode schizophrenia patients before and after 6 months' treatment with quetiapine. *Int. J. Neuropsychopharmacol.*, 13(10):1383–1395.

[Ambrosen, 2017] Ambrosen, K. M. S. (2017). *Modeling Structural Brain Connectivity*. PhD thesis, Technical University of Denmark (DTU).

[American Psychiatric Association, 2013] American Psychiatric Association (2013). *Diagnostic and Statistical Manual of Mental Disorders*. DSM Library. American Psychiatric Association.

[Andersen et al., 2011] Andersen, R., Fagerlund, B., Rasmussen, H., Ebdrup, B. H., Aggernaes, B., Gade, A., Oranje, B., and Glenthoj, B. (2011). Cognitive effects of six months of treatment with quetiapine in antipsychotic-naïve first-episode schizophrenia. *Psychiatry Res.*, 187(1-2):49–54.

[Andreasen et al., 2005] Andreasen, N. C., Carpenter, W. T., Kane, J. M., Lasser, R. A., Marder, S. R., and Weinberger, D. R. (2005). Remission in schizophrenia: Proposed criteria and rationale for consensus. *Am. J. Psychiatry*, 162(3):441–449.

[Arbabshirani et al., 2017] Arbabshirani, M. R., Plis, S., Sui, J., and Calhoun, V. D. (2017). Single subject prediction of brain disorders in neuroimaging: Promises and pitfalls. *Neuroimage*, 145(Pt B):137–165.

[Atrey et al., 2010] Atrey, P. K., Hossain, M. A., El Saddik, A., and Kankanhalli, M. S. (2010). Multimodal fusion for multimedia analysis: A survey. *Multimedia Systems*, 16:345–379.

[Braff et al., 2001] Braff, D. L., Geyer, M. A., and Swerdlow, N. R. (2001). Human studies of prepulse inhibition of startle: normal subjects, patient groups, and pharmacological studies. *Psychopharmacology*, 156(2):234–258.

[Bzdok and Meyer-Lindenberg, 2018] Bzdok, D. and Meyer-Lindenberg, A. (2018). Machine learning for precision psychiatry: Opportunities and challenges. *Biol Psychiatry Cogn Neurosci Neuroimaging*, 3(3):223–230.

[Calhoun and Sui, 2016] Calhoun, V. D. and Sui, J. (2016). Multimodal fusion of brain imaging data: A key to finding the missing link(s) in complex mental illness. *Biol Psychiatry Cogn Neurosci Neuroimaging*, 1(3):230–244.

[Charney et al., 2017] Charney, A. W., Ruderfer, D. M., Stahl, E. A., Moran, J. L., Chambert, K., Belliveau, R. A., Forty, L., Gordon-Smith, K., Di Florio, A., Lee, P. H., Bromet, E. J., Buckley, P. F., Escamilla, M. A., Fanous, A. H., Fochtmann, L. J., Lehrer, D. S., Malaspina, D., Marder, S. R., Morley, C. P., Nicolini, H., Perkins, D. O., Rakofsky, J. J., Rapaport, M. H., Medeiros, H., Sobell, J. L., Green, E. K., Backlund, L., Bergen, S. E., Juréus, A., Schalling, M., Lichtenstein, P., Roussos, P., Knowles, J. A., Jones, I., Jones, L. A., Hultman, C. M., Perlis, R. H., Purcell, S. M., McCarroll, S. A., Pato, C. N., Pato, M. T., Craddock, N., Landén, M., Smoller, J. W., and Sklar, P. (2017). Evidence for genetic heterogeneity between clinical subtypes of bipolar disorder. *Transl. Psychiatry*, 7(1):e993.

[Chawla et al., 2002] Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *J. Artif. Intell. Res.*, 16:321–357.

[Cuthbert and Insel, 2013] Cuthbert, B. N. and Insel, T. R. (2013). Toward the future of psychiatric diagnosis: the seven pillars of RDoC. *BMC Med.*, 11:126.

[Dähne et al., 2015] Dähne, S., Bießmann, F., Samek, W., Haufe, S., Goltz, D., Gundlach, C., Villringer, A., Fazli, S., and Müller, K. (2015). Multivariate machine learning methods for fusing multimodal functional neuroimaging data. *Proc. IEEE*, 103(9):1507–1530.

[Deng et al., 2009] Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *In CVPR*.

[Desikan et al., 2006] Desikan, R. S., Ségonne, F., Fischl, B., Quinn, B. T., Dickerson, B. C., Blacker, D., Buckner, R. L., Dale, A. M., Maguire, R. P., Hyman, B. T., Albert, M. S., and Killiany, R. J. (2006). An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. *Neuroimage*, 31(3):968–980.

[DeVries and Taylor, 2017] DeVries, T. and Taylor, G. W. (2017). Dataset augmentation in feature space.

[Dheeru and Karra Taniskidou, 2017] Dheeru, D. and Karra Taniskidou, E. (2017). UCI machine learning repository. `http://archive.ics.uci.edu/ml`.

[Domingos and Pazzani, 1997] Domingos, P. and Pazzani, M. (1997). On the optimality of the simple bayesian classifier under Zero-One loss. *Mach. Learn.*, 29(2):103–130.

[Donders et al., 2006] Donders, A. R. T., van der Heijden, G. J. M. G., Stijnen, T., and Moons, K. G. M. (2006). Review: a gentle introduction to imputation of missing values. *J. Clin. Epidemiol.*, 59(10):1087–1091.

[Düring et al., 2014] Düring, S., Glenthøj, B. Y., Andersen, G. S., and Oranje, B. (2014). Effects of dopamine D2/D3 blockade on human sensory and sensorimotor gating in initially antipsychotic-naive, first-episode schizophrenia patients. *Neuropsychopharmacology*, 39(13):3000–3008.

[Düring et al., 2015] Düring, S., Glenthøj, B. Y., and Oranje, B. (2015). Effects of blocking D2/D3 receptors on mismatch negativity and p3a amplitude of initially antipsychotic naïve, first episode schizophrenia patients. *Int. J. Neuropsychopharmacol.*, 19(3):yv109.

[Ebdrup, 2009] Ebdrup, B. H. (2009). *Structual brain changes in antipsychotic-naïve first-episode schizophrenia patients before and after six months of antipsychotic monotherapy.* PhD thesis, Copenhagen University Hospital Glostrup.

[Ebdrup et al., 2010] Ebdrup, B. H., Glenthøj, B., Rasmussen, H., Aggernaes, B., Langkilde, A. R., Paulson, O. B., Lublin, H., Skimminge, A., and Baaré, W. (2010). Hippocampal and caudate volume reductions in antipsychotic-naive first-episode schizophrenia. *J. Psychiatry Neurosci.*, 35(2):95–104.

[Ebdrup et al., 2016] Ebdrup, B. H., Raghava, J. M., Nielsen, M. Ø., Rostrup, E., and Glenthøj, B. (2016). Frontal fasciculi and psychotic symptoms in antipsychotic-naive patients with schizophrenia before and after 6 weeks of selective dopamine d2/3 receptor blockade. *J. Psychiatry Neurosci.*, 41(2):133–141.

[Fagerlund et al., 2004] Fagerlund, B., Mackeprang, T., Gade, A., and Glenthøj, B. Y. (2004). Effects of low-dose risperidone and low-dose zuclopenthixol on cognitive functions in first-episode drug-naive schizophrenic patients. *CNS Spectr.*, 9(5):364–374.

[Fusar-Poli et al., 2016] Fusar-Poli, P., Cappucciati, M., Rutigliano, G., Heslin, M., Stahl, D., Brittenden, Z., Caverzasi, E., McGuire, P., and Carpenter, W. T. (2016). Diagnostic stability of ICD/DSM first episode psychosis diagnoses: Meta-analysis. *Schizophr. Bull.*, 42(6):1395–1406.

[Fusar-Poli et al., 2013] Fusar-Poli, P., Smieskova, R., Kempton, M. J., Ho, B. C., Andreasen, N. C., and Borgwardt, S. (2013). Progressive brain changes in schizophrenia related to antipsychotic treatment? a meta-analysis of longitudinal MRI studies. *Neurosci. Biobehav. Rev.*, 37(8):1680–1691.

[Gandal et al., 2018] Gandal, M. J., Haney, J. R., Parikshak, N. N., Leppa, V., Ramaswami, G., Hartl, C., Schork, A. J., Appadurai, V., Buil, A., Werge, T. M., Liu, C., White, K. P., CommonMind Consortium, PsychENCODE Consortium, iPSYCH-BROAD Working Group, Horvath, S., and Geschwind, D. H. (2018). Shared molecular neuropathology across major psychiatric disorders parallels polygenic overlap. *Science*, 359(6376):693–697.

[Goodkind et al., 2015] Goodkind, M., Eickhoff, S. B., Oathes, D. J., Jiang, Y., Chang, A., Jones-Hagata, L. B., Ortega, B. N., Zaiko, Y. V., Roach, E. L., Korgaonkar, M. S., Grieve, S. M., Galatzer-Levy, I., Fox, P. T., and Etkin, A. (2015). Identification of a common neurobiological substrate for mental illness. *JAMA Psychiatry*, 72(4):305–315.

[Gottesman and Gould, 2003] Gottesman, I. I. and Gould, T. D. (2003). Reviews and overviews the endophenotype concept in psychiatry : Etymology and strategic intentions. *Am. J. Psychiatry*, (April):636–645.

[Haigh et al., 2017] Haigh, S. M., Coffman, B. A., and Salisbury, D. F. (2017). Mismatch negativity in First-Episode schizophrenia: A Meta-Analysis. *Clin. EEG Neurosci.*, 48(1):3–10.

[Haijma et al., 2013] Haijma, S. V., Van Haren, N., Cahn, W., Koolschijn, P. C. M. P., Hulshoff Pol, H. E., and Kahn, R. S. (2013). Brain volumes in schizophrenia: a meta-analysis in over 18 000 subjects. *Schizophr. Bull.*, 39(5):1129–1138.

[Hall and Llinas, 1997] Hall, D. L. D. L. and Llinas, J. (1997). An introduction to multisensor data fusion. *Proc. IEEE*, 85(1):6–23.

[Hansen et al., 1999] Hansen, L. K., Larsen, J., Nielsen, F. A., Strother, S. C., Rostrup, E., Savoy, R., Lange, N., Sidtis, J., Svarer, C., and Paulson, O. B. (1999). Generalizable patterns in neuroimaging: how many principal components? *Neuroimage*, 9(5):534–544.

[Hansen et al., 2015] Hansen, S. T., Winkler, I., Hansen, L. K., Muller, K. R., and others (2015). Fusing simultaneous EEG and fMRI using functional and anatomical information. *Workshop on Pattern ...*.

[Hanson, 2009] Hanson, L. G. (2009). Introduction to magnetic resonance imaging techniques. Clinical and research applications of diagnostic imaging techniques: MR, PET, SPECT, CT and ultrasound: PhD Course.

[Hastie et al., 2009] Hastie, T., Tibshirani, R. J., and Friedman, J. (2009). *The Elements of Statistical Learning*, volume 1. Springer, second edition.

[Hauberg et al., 2016] Hauberg, S., Freifeld, O., Larsen, A. B. L., Fisher, J., and Hansen, L. (2016). Dreaming more data: Class-dependent distributions over diffeomorphisms for learned data augmentation. In *Artificial Intelligence and Statistics*, volume 51, pages 342–350. jmlr.org.

[Hedman et al., 2013] Hedman, A. M., van Haren, N. E. M., van Baal, C. G. M., Kahn, R. S., and Hulshoff Pol, H. E. (2013). IQ change over time in schizophrenia and healthy individuals: a meta-analysis. *Schizophr. Res.*, 146(1-3):201–208.

[Hinton, 2002] Hinton, G. E. (2002). Training products of experts by minimizing contrastive divergence. *Neural Comput.*, 14:1771–1800.

[Hotelling, 1936] Hotelling, H. (1936). Relations between two sets of variates. *Biometrika*, 28(3/4):321–377.

[Howes et al., 2015] Howes, O., McCutcheon, R., and Stone, J. (2015). Glutamate and dopamine in schizophrenia: an update for the 21st century. *J. Psychopharmacol.*, 29(2):97–115.

[Huys et al., 2016] Huys, Q. J. M., Maia, T. V., and Frank, M. J. (2016). Computational psychiatry as a bridge from neuroscience to clinical applications. *Nat. Neurosci.*, 19(3):404–413.

[Insel, 2013] Insel, T. R. (2013). Transforming diagnosis. `https://www.nimh.nih.gov/about/directors/thomas-insel/blog/2013/transforming-diagnosis.shtml`. Accessed: 2018-11-11.

[Itti and Baldi, 2009] Itti, L. and Baldi, P. (2009). Bayesian surprise attracts human attention. *Vision Res.*, 49(10):1295–1306.

[Jakobsen et al., 2005] Jakobsen, K. D., Frederiksen, J. N., Hansen, T., Jansson, L. B., Parnas, J., and Werge, T. (2005). Reliability of clinical ICD-10 schizophrenia diagnoses. *Nord. J. Psychiatry*, 59(3):209–212.

[Jenkinson et al., 2012] Jenkinson, M., Beckmann, C. F., Behrens, T. E. J., Woolrich, M. W., and Smith, S. M. (2012). FSL. *Neuroimage*, 62(2):782–790.

[Jessen et al., 2018] Jessen, K., Rostrup, E., Mandl, R. C. W., Nielsen, M. Ø., Bak, N., Fagerlund, B., Glenthøj, B. Y., and Ebdrup, B. H. (2018). Cortical structures and their clinical correlates in antipsychotic-naïve schizophrenia patients before and after 6 weeks of dopamine d2/3 receptor antagonist treatment. *Psychol. Med.*, pages 1–10.

[Kahn and Keefe, 2013] Kahn, R. S. and Keefe, R. S. E. (2013). Schizophrenia is a cognitive illness. *JAMA Psychiatry*, 70(10):1107.

[Kambeitz et al., 2015] Kambeitz, J., Kambeitz-Ilankovic, L., Leucht, S., Wood, S., Davatzikos, C., Malchow, B., Falkai, P., and Koutsouleris, N. (2015). Detecting neuroimaging biomarkers for schizophrenia: a meta-analysis of multivariate pattern recognition studies. *Neuropsychopharmacology*, 40(7):1742–1751.

[Kapur, 2011] Kapur, S. (2011). What kraepelin might say about schizophrenia: just the facts. *Schizophr. Res.*, 128(1-3):1–2.

[Kelly et al., 2018] Kelly, S., Jahanshad, N., Zalesky, A., Kochunov, P., Agartz, I., Alloza, C., Andreassen, O. A., Arango, C., Banaj, N., Bouix, S., Bousman, C. A., Brouwer, R. M., Bruggemann, J., Bustillo, J., Cahn, W., Calhoun, V., Cannon, D., Carr, V., Catts, S., Chen, J., Chen, J.-X., Chen, X., Chiapponi, C., Cho, K. K., Ciullo, V., Corvin, A. S., Crespo-Facorro, B., Cropley, V., De Rossi, P., Diaz-Caneja, C. M., Dickie, E. W., Ehrlich, S., Fan, F.-M., Faskowitz, J., Fatouros-Bergman, H., Flyckt, L., Ford, J. M., Fouche, J.-P., Fukunaga, M., Gill, M., Glahn, D. C., Gollub, R., Goudzwaard, E. D., Guo, H., Gur, R. E., Gur, R. C., Gurholt, T. P., Hashimoto, R., Hatton, S. N., Henskens, F. A., Hibar, D. P., Hickie, I. B., Hong, L. E., Horacek, J., Howells, F. M., Hulshoff Pol, H. E., Hyde, C. L., Isaev, D., Jablensky, A., Jansen, P. R., Janssen, J., Jönsson, E. G., Jung, L. A., Kahn, R. S., Kikinis, Z., Liu, K., Klauser, P., Knöchel, C., Kubicki, M., Lagopoulos, J., Langen, C., Lawrie, S., Lenroot, R. K., Lim, K. O., Lopez-Jaramillo, C., Lyall, A., Magnotta, V., Mandl, R. C. W., Mathalon, D. H., McCarley, R. W., McCarthy-Jones, S., McDonald, C., McEwen, S., McIntosh, A., Melicher, T., Mesholam-Gately, R. I., Michie, P. T., Mowry, B., Mueller, B. A., Newell, D. T., O'Donnell, P., Oertel-Knöchel, V., Oestreich, L., Paciga, S. A., Pantelis, C., Pasternak, O., Pearlson, G., Pellicano, G. R., Pereira, A., Pineda Zapata, J., Piras, F., Potkin, S. G., Preda, A., Rasser, P. E., Roalf, D. R., Roiz, R., Roos, A., Rotenberg, D., Satterthwaite, T. D., Savadjiev, P., Schall, U., Scott, R. J., Seal, M. L., Seidman, L. J., Shannon Weickert, C., Whelan, C. D., Shenton, M. E., Kwon, J. S., Spalletta, G., Spaniel, F., Sprooten, E., Stäblein, M., Stein, D. J., Sundram, S., Tan, Y., Tan, S., Tang, S., Temmingh, H. S., Westlye, L. T., Tønnesen, S., Tordesillas-Gutierrez, D., Doan, N. T., Vaidya, J., van Haren, N. E. M., Vargas, C. D., Vecchio, D., Velakoulis, D., Voineskos, A., Voyvodic, J. Q., Wang, Z., Wan, P., Wei, D., Weickert, T. W., Whalley,

H., White, T., Whitford, T. J., Wojcik, J. D., Xiang, H., Xie, Z., Yamamori, H., Yang, F., Yao, N., Zhang, G., Zhao, J., van Erp, T. G. M., Turner, J., Thompson, P. M., and Donohoe, G. (2018). Widespread white matter microstructural differences in schizophrenia across 4322 individuals: results from the ENIGMA schizophrenia DTI working group. *Mol. Psychiatry*, 23(5):1261–1269.

[Keshavan et al., 2008] Keshavan, M. S., Tandon, R., Boutros, N. N., and Nasrallah, H. A. (2008). Schizophrenia, "just the facts": what we know in 2008 part 3: neurobiology. *Schizophr. Res.*, 106(2-3):89–107.

[Khaleghi et al., 2013] Khaleghi, B., Khamis, A., Karray, F. O., and Razavi, S. N. (2013). Multisensor data fusion: A review of the state-of-the-art. *Inf. Fusion*, 14(1):28–44.

[Kittler et al., 1998] Kittler, J., Hatef, M., Duin, R. P. W., and Matas, J. (1998). On combining classifiers. *IEEE Trans. Pattern Anal. Mach. Intell.*, 20(3):226–239.

[Krizhevsky et al., 2012] Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. In Pereira, F., Burges, C. J. C., Bottou, L., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc.

[Lahat et al., 2015] Lahat, D., Adali, T., and Jutten, C. (2015). Multimodal data fusion: An overview of methods, challenges, and prospects. *Proc. IEEE*, 103(9):1449–1477.

[Laruelle et al., 1996] Laruelle, M., Abi-Dargham, A., van Dyck, C. H., Gil, R., D'Souza, C. D., Erdos, J., McCance, E., Rosenblatt, W., Fingado, C., Zoghbi, S. S., Baldwin, R. M., Seibyl, J. P., Krystal, J. H., Charney, D. S., and Innis, R. B. (1996). Single photon emission computerized tomography imaging of amphetamine-induced dopamine release in drug-free schizophrenic subjects. *Proc. Natl. Acad. Sci. U. S. A.*, 93(17):9235–9240.

[Lecun et al., 1998] Lecun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proc. IEEE*, 86(11):2278–2324.

[Li et al., 2014] Li, R., Zhang, W., Suk, H.-I., Wang, L., Li, J., Shen, D., and Ji, S. (2014). Deep learning based imaging data completion for improved brain disease diagnosis. *Med. Image Comput. Comput. Assist. Interv.*, 17(Pt 3):305–312.

[Mackeprang et al., 2002] Mackeprang, T., Kristiansen, K. T., and Glenthoj, B. Y. (2002). Effects of antipsychotics on prepulse inhibition of the startle

response in drug-naïve schizophrenic patients. *Biol. Psychiatry*, 52(9):863–873.

[Masi et al., 2017] Masi, A., DeMayo, M. M., Glozier, N., and Guastella, A. J. (2017). An overview of autism spectrum disorder, heterogeneity and treatment options. *Neurosci. Bull.*, 33(2):183–193.

[McGrath et al., 2008] McGrath, J., Saha, S., Chant, D., and Welham, J. (2008). Schizophrenia: a concise overview of incidence, prevalence, and mortality. *Epidemiol. Rev.*, 30:67–76.

[Monteiro et al., 2016] Monteiro, J. M., Rao, A., Shawe-Taylor, J., Mourão-Miranda, J., and Alzheimer's Disease Initiative (2016). A multiple hold-out framework for sparse partial least squares. *J. Neurosci. Methods*, 271:182–194.

[Ng and Jordan, 2002] Ng, A. Y. and Jordan, M. I. (2002). On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. In Dietterich, T. G., Becker, S., and Ghahramani, Z., editors, *Advances in neural information processing systems*, pages 841–848. MIT Press.

[Nørbak-Emig et al., 2017] Nørbak-Emig, H., Pinborg, L. H., Raghava, J. M., Svarer, C., Baaré, W. F. C., Allerup, P., Friberg, L., Rostrup, E., Glenthøj, B., and Ebdrup, B. H. (2017). Extrastriatal dopamine d2/3 receptors and cortical grey matter volumes in antipsychotic-naïve schizophrenia patients before and after initial antipsychotic treatment. *World J. Biol. Psychiatry*, 18(7):539–549.

[Nuechterlein et al., 2004] Nuechterlein, K. H., Barch, D. M., Gold, J. M., Goldberg, T. E., Green, M. F., and Heaton, R. K. (2004). Identification of separable cognitive factors in schizophrenia. *Schizophr. Res.*, 72(1):29–39.

[Oranje et al., 2013] Oranje, B., Aggernaes, B., Rasmussen, H., Ebdrup, B. H., and Glenthøj, B. Y. (2013). P50 suppression and its neural generators in antipsychotic-naive first-episode schizophrenia before and after 6 months of quetiapine treatment. *Schizophr. Bull.*, 39(2):472–480.

[Oranje et al., 2017] Oranje, B., Aggernaes, B., Rasmussen, H., Ebdrup, B. H., and Glenthøj, B. Y. (2017). Selective attention and mismatch negativity in antipsychotic-naïve, first-episode schizophrenia patients before and after 6 months of antipsychotic monotherapy. *Psychol. Med.*, 47(12):2155–2165.

[Orrù et al., 2012] Orrù, G., Pettersson-Yeo, W., Marquand, A. F., Sartori, G., and Mechelli, A. (2012). Using support vector machine to identify imaging biomarkers of neurological and psychiatric disease: A critical review. *Neurosci. Biobehav. Rev.*, 36(4):1140–1152.

[Owens et al., 2016] Owens, E. M., Bachman, P., Glahn, D. C., and Bearden, C. E. (2016). Electrophysiological endophenotypes for schizophrenia. *Harv. Rev. Psychiatry*, 24(2):129–147.

[Pearl, 1988] Pearl, J. (1988). On probability intervals. *Int. J. Approx. Reason.*, 2(3):211–216.

[Pohl and Van Genderen, 1998] Pohl, C. and Van Genderen, J. L. (1998). Review article multisensor image fusion in remote sensing: Concepts, methods and applications. *Int. J. Remote Sens.*, 19:823–854.

[Prata et al., 2014] Prata, D., Mechelli, A., and Kapur, S. (2014). Clinically meaningful biomarkers for psychosis: a systematic and quantitative review. *Neurosci. Biobehav. Rev.*, 45:134–141.

[Robbins et al., 1994] Robbins, T. W., James, M., Owen, A. M., Sahakian, B. J., McInnes, L., and Rabbitt, P. (1994). Cambridge neuropsychological test automated battery (CANTAB): a factor analytic study of a large sample of normal elderly volunteers. *Dementia*, 5(5):266–281.

[Schaefer et al., 2013] Schaefer, J., Giangrande, E., Weinberger, D. R., and Dickinson, D. (2013). The global cognitive impairment in schizophrenia: consistent over decades and around the world. *Schizophr. Res.*, 150(1):42–50.

[Smolensky, 1986] Smolensky, P. (1986). Information processing in dynamical systems: Foundations of harmony theory. In Rumelhart, D. E. and McClelland, J. L., editor, *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, volume 1, chapter 6, pages 194–281. dtic.mil.

[Sui et al., 2012] Sui, J., Adali, T., Yu, Q., Chen, J., and Calhoun, V. D. (2012). A review of multivariate methods for multimodal fusion of brain imaging data. *J. Neurosci. Methods*, 204(1):68–81.

[Tandon et al., 2008] Tandon, R., Keshavan, M. S., and Nasrallah, H. A. (2008). Schizophrenia, "just the facts" what we know in 2008. 2. epidemiology and etiology. *Schizophr. Res.*, 102(1-3):1–18.

[Tandon et al., 2009] Tandon, R., Nasrallah, H. A., and Keshavan, M. S. (2009). Schizophrenia, "just the facts" 4. clinical features and conceptualization. *Schizophr. Res.*, 110(1-3):1–23.

[Tandon et al., 2010] Tandon, R., Nasrallah, H. A., and Keshavan, M. S. (2010). Schizophrenia, "just the facts" 5. treatment and prevention. past, present, and future. *Schizophr. Res.*, 122(1-3):1–23.

[van Erp et al., 2018] van Erp, T. G. M., Walton, E., Hibar, D. P., Schmaal, L., Jiang, W., Glahn, D. C., Pearlson, G. D., Yao, N., Fukunaga, M., Hashimoto, R., Okada, N., Yamamori, H., Bustillo, J. R., Clark, V. P., Agartz, I.,

Mueller, B. A., Cahn, W., de Zwarte, S. M. C., Hulshoff Pol, H. E., Kahn, R. S., Ophoff, R. A., van Haren, N. E. M., Andreassen, O. A., Dale, A. M., Doan, N. T., Gurholt, T. P., Hartberg, C. B., Haukvik, U. K., Jørgensen, K. N., Lagerberg, T. V., Melle, I., Westlye, L. T., Gruber, O., Kraemer, B., Richter, A., Zilles, D., Calhoun, V. D., Crespo-Facorro, B., Roiz-Santiañez, R., Tordesillas-Gutiérrez, D., Loughland, C., Carr, V. J., Catts, S., Cropley, V. L., Fullerton, J. M., Green, M. J., Henskens, F. A., Jablensky, A., Lenroot, R. K., Mowry, B. J., Michie, P. T., Pantelis, C., Quidé, Y., Schall, U., Scott, R. J., Cairns, M. J., Seal, M., Tooney, P. A., Rasser, P. E., Cooper, G., Shannon Weickert, C., Weickert, T. W., Morris, D. W., Hong, E., Kochunov, P., Beard, L. M., Gur, R. E., Gur, R. C., Satterthwaite, T. D., Wolf, D. H., Belger, A., Brown, G. G., Ford, J. M., Macciardi, F., Mathalon, D. H., O'Leary, D. S., Potkin, S. G., Preda, A., Voyvodic, J., Lim, K. O., McEwen, S., Yang, F., Tan, Y., Tan, S., Wang, Z., Fan, F., Chen, J., Xiang, H., Tang, S., Guo, H., Wan, P., Wei, D., Bockholt, H. J., Ehrlich, S., Wolthusen, R. P. F., King, M. D., Shoemaker, J. M., Sponheim, S. R., De Haan, L., Koenders, L., Machielsen, M. W., van Amelsvoort, T., Veltman, D. J., Assogna, F., Banaj, N., de Rossi, P., Iorio, M., Piras, F., Spalletta, G., McKenna, P. J., Pomarol-Clotet, E., Salvador, R., Corvin, A., Donohoe, G., Kelly, S., Whelan, C. D., Dickie, E. W., Rotenberg, D., Voineskos, A. N., Ciufolini, S., Radua, J., Dazzan, P., Murray, R., Reis Marques, T., Simmons, A., Borgwardt, S., Egloff, L., Harrisberger, F., Riecher-Rössler, A., Smieskova, R., Alpert, K. I., Wang, L., Jönsson, E. G., Koops, S., Sommer, I. E. C., Bertolino, A., Bonvino, A., Di Giorgio, A., Neilson, E., Mayer, A. R., Stephen, J. M., Kwon, J. S., Yun, J.-Y., Cannon, D. M., McDonald, C., Lebedeva, I., Tomyshev, A. S., Akhadov, T., Kaleda, V., Fatouros-Bergman, H., Flyckt, L., Karolinska Schizophrenia Project, Busatto, G. F., Rosa, P. G. P., Serpa, M. H., Zanetti, M. V., Hoschl, C., Skoch, A., Spaniel, F., Tomecek, D., Hagenaars, S. P., McIntosh, A. M., Whalley, H. C., Lawrie, S. M., Knöchel, C., Oertel-Knöchel, V., Stäblein, M., Howells, F. M., Stein, D. J., Temmingh, H. S., Uhlmann, A., Lopez-Jaramillo, C., Dima, D., McMahon, A., Faskowitz, J. I., Gutman, B. A., Jahanshad, N., Thompson, P. M., and Turner, J. A. (2018). Cortical brain abnormalities in 4474 individuals with schizophrenia and 5098 control subjects via the enhancing neuro imaging genetics through meta analysis (ENIGMA) consortium. *Biol. Psychiatry*, 84(9):644–654.

[Wolfers et al., 2015] Wolfers, T., Buitelaar, J. K., Beckmann, C. F., Franke, B., and Marquand, A. F. (2015). From estimating activation locality to predicting disorder: A review of pattern recognition for neuroimaging-based psychiatric diagnostics. *Neurosci. Biobehav. Rev.*, 57:328–349.

[World Health Organization, 2018] World Health Organization (2018). *International statistical classification of diseases and related health problems (11th Revision).*

[Xiang et al., 2013] Xiang, S., Yuan, L., Fan, W., Wang, Y., Thompson, P. M., and Ye, J. (2013). Bi-level multi-source learning for heterogeneous block-wise missing data. *Neuroimage*, 102:192–206.

[Xu et al., 1992] Xu, L., Krzyzak, A., and Suen, C. Y. (1992). Methods of combining multiple classifiers and their applications to handwriting recognition. *IEEE Trans. Syst. Man Cybern.*, 22(3):418–435.

[Xue and Titterington, 2008] Xue, J. H. and Titterington, D. M. (2008). Comment on "on discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes". *Neural Process. Letters*, 28(3):169–187.

[Yuan et al., 2012] Yuan, L., Wang, Y., Thompson, P. M., Narayan, V. A., Ye, J., and Alzheimer's Disease Neuroimaging Initiative (2012). Multi-source feature learning for joint analysis of incomplete multiple heterogeneous neuroimaging data. *Neuroimage*, 61(3):622–632.

[Zhang, 2004] Zhang, H. (2004). The optimality of naive bayes. *Archit. Aujourdhui.*, 1(2):3.

[Zhao et al., 2017] Zhao, J., Xie, X., Xu, X., and Sun, S. (2017). Multi-view learning overview: Recent progress and new challenges. *Inf. Fusion*, 38:43–54.

[Zhu et al., 2018] Zhu, H., Li, G., and Lock, E. F. (2018). Generalized integrative principal component analysis for multi-type data with block-wise missing structure. *Biostatistics*.